

算法统治世界

智能经济的隐形秩序

徐恪　李沁　著

The Invisible Order of
the Intelligent Economy

畅销书《赛博新经济》作者最新力作

清华大学出版社

算法统治世界

智能经济的隐形秩序

徐恪　李沁　著

清 华 大 学 出 版 社
北　京

内 容 简 介

今天，互联网已经彻底改变了经济系统的运行方式，经济增长的决定性要素已经从物质资料的增加转变成为信息的增长。但是，只有信息的快速增长是不够的，这些增长的信息还必须是“有序”的。只有“有序”才能使信息具有价值，能够为人所用，能够指导我们实现商业的新路径。这种包含在信息里的隐形秩序才是今天信息世界的真正价值所在。经济系统的运行确实是纷繁复杂的，但因为算法的存在，这一切变得有律可循，算法也成为新经济系统里那只“看不见的手”。那么算法究竟是如何保障信息的有序增长的？又是如何改变和控制我们的经济系统的？在这本书中，我们将赛博经济系统从上至下进行了梳理，把推荐算法、分配算法、匹配算法、动态定价算法、大数据处理算法、数据交易算法、隐私保护算法和区块链技术及相关算法在新经济组织中的运行做了深入浅出的阐述，从而为读者打开了新经济的现在与未来的大门。本书虽有专业的深度，但也适合一般读者阅读，同时这些创新背后的算法逻辑也将帮助企业更好地规划自己的商业模式和未来战略。

图书在版编目(CIP)数据

算法统治世界：智能经济的隐形秩序/徐恪，李沁著. —北京：清华大学出版社，2017(2017.12 重印)
ISBN 978-7-302-48808-8

Ⅰ. ①算…　Ⅱ. ①徐…　②李…　Ⅲ. ①计算机算法—应用—经济系统—研究　Ⅳ. ①F014.9

中国版本图书馆 CIP 数据核字(2017)第 265983 号

责任编辑：龙启铭
封面设计：傅瑞学
责任校对：李建庄
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编**：100084
社 总 机：010-62770175　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京亿浓世纪彩色印刷有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　**印张**：23.5　**插页**：1　**字　　数**：442 千字
版　　次：2017 年 12 月第 1 版　**印　　次**：2017 年 12 月第2 次印刷
印　　数：3001～5000
定　　价：69.00 元

产品编号：076220-02

自序

如今,互联网如同高速公路一样,成为了人类社会的信息基础设施,在经济领域催生出众多前所未有的商业新形态,并开始影响和改变整体社会经济系统的运行方式。当网络与经济系统不断融合,催生出了一种全新的经济系统,我们将之称为“赛博经济系统”(Cyber Economic System),它是以广泛连接和海量数据为基础,以互联网技术为载体,以算法为内在驱动力的新型经济系统。

新经济系统已经到来,但我们还没来得及真正深入探寻其内在规律,特别是2008年的经济危机,很少有经济学家做出提前预测的现实,让经济学领域一时陷入低谷,纳西姆·塔勒布在《黑天鹅》一书中将原因归结为“因变化而需要知识的事物,通常是没有专家的”。的确,经济学面对的问题太复杂,复杂到几乎不可预测。虽然经济领域正变

得越来越复杂，但天气系统其实同样复杂，而借助超级计算机，我们至少可以预报近期的天气。类似的例子还有AlphoGo，当AlphoGo战胜李世石时，人类棋手第一人柯洁痛苦地承认："人类几千年甚至没人沾到围棋真理的边"，而AlphoGo为围棋真理打开了新的大门。这些例子在启发我们思考，是不是我们也没有沾到经济系统运行真理的边？也许我们需要一些新的思路。

麻省理工学院的物理学家塞萨尔·伊达尔戈在《增长的本质》一书中给出了一个有趣的想法："经济是通过实体化信息，增强人类对于知识的实际应用的系统"，"经济增长的本质，是因为信息(量)的增长"。如果我们采用"增长"的概念来重新定义赛博新经济，那么赛博新经济是信息增长更快的经济。赛博世界为信息的增长提供了快速路径，这是由赛博的特征决定的。赛博的特征包括小世界、信息级联、幂律和长尾等，这些都是赛博世界所独有的，这也让赛博时代的信息增长速度超过以往任何时候。

不过，只有信息的快速增长，还是不够的。这些增长的信息还必须是"有序"的。只有"有序"才能使信息具有价值，能够为人所用，这是一种隐形的秩序。这种包含在信息里的隐形秩序，才是今天信息世界的真正价值所在。那么我们该如何保证信息按照有序的方式增长？经济系统的运行确实是纷繁复杂的，但因为算法的存在，使一切变得有律可循，从某种意义上说，算法已经成为新经济系统里那只"看不见的手"。

那么算法究竟是什么？又有哪些算法正在改变和控制着我们的经济系统？在这本书中，我们将赛博经济系统从上至下进行了梳理，把推荐算法、分配算法、匹配算法、动态定价算法、区块链技术及相关算法、大数据处理算法、数据交易算法和隐私保护算法如何推动新经济系统的运行做了详细阐述。当然，仅仅理解和掌握这些算法，并不能保证对经济系统运行做出准确预测，但当我们理解了这些算法和它们所代表的赛博经济的隐形秩序后，或许再次面对经济系统的不确定性时，会增加一份信心，相信也会帮助大家从中找到

解决问题的全新视角。

这里要多说一句，身为计算机科学工作者，从事经济学方面的研究似乎有不务正业之嫌，但最近经济学知名期刊 *Econometrica* 的主编、经济学家 Joel Sobel 在一次会议报告中指出，当前经济学中最有趣的研究是计算机科学家完成的，因为计算机科学家有场景、有平台并能真正通过实验验证理论。这给了我这样的研究者莫大的鼓励，我个人认为，未来经济学的重大突破很可能会来自计算机科学工作者和经济学工作者的合作研究，让我们共同努力，使这一天早日到来。

本书的研究得到了国家自然科学基金(61472212)支持，在此深表谢意。在本书的编写过程中，我们借鉴吸收了许多国内外专家学者的研究成果，在此也致以诚挚的谢意。本书部分内容曾经在“网络科学与策略机制”“互联网发展导论”和“互联网发展与创新经济”课程中进行过讲授，同学们也提出了非常宝贵的意见，在此一并致谢。教学的快乐也许就在于此吧。感谢我的研究生苏辉、李立、李彤、吴波、吕亮、付瑶、张欣欣、张宇超、杨帆、姚文兵，感谢我的合作者李沁女士。还要感谢清华大学社科学院的王勇教授，他的洞见和观察使我获益良多。

限于水平，书中不可避免存在欠缺之处，恳请读者批评指正。读者可以通过关注“赛博新经济”公众号和我们交流。

徐恪

2017年 10月

前言

算法定义的新经济系统

5G通信技术已经悄然出现,这个号称“人类历史上最复杂的通信系统”,预计将在2020年开始商用化;2016年3月15日,当韩国棋手李世石在第180手投子认输的那一刹那,宣告了以AlphaGo为代表的人工智能技术开启了崭新的一页;2016年8月,中国将世界上第一颗量子通信卫星“墨子号”送入太空轨道,并完成了人类历史上首次量子通信过程;中国在关键技术上的重大突破,有望使量子计算机在四五年内进入实用阶段……人类正在加速进入赛博时代。

如果把时间倒推三十多年[①],那时的人们应该很难想象:用来打电话的手机能够完成这么多事情;素不相识人

① 1986年,美国NSFNET建成,有人认为这是互联网真正诞生的标志性事件。

之间可以跨越时间和地域的限制相识相知；任何一个角落里发生的事情都可以在几分钟之内传遍世界；普通人也能一呼百应，影响千千万万的人。赛博时代，高度发达的互联网络在人与人之间、人与物之间，架起一道道看不见的连接线，将素不相识的人、貌似毫无关联的物，连接在了一起，缩短了时空的距离。

同时，这种无所不连的连接也打破了原有经济系统中的生产、消费、市场等环节，形成了覆盖更广、连接度更高、效率更快的新经济系统，即赛博新经济。赛博新经济自出生之日起就展现出极强的生命力，发展速度远超过去任何一个经济时代，这其中还涌现出许多充满活力的新机制、新模式。随着算法不断地渗透，信息增长的秩序和赛博新经济系统的运行秩序也不断地被它重新定义和改写，从某种程度上，我们甚至可以说，它决定了经济系统的不断进化。

信息与秩序

人们曾经以采集食物为生，而如今他们要重新以采集信息为生，尽管这件事看起来很不可思议。

——马歇尔·麦克卢汉，原创媒介理论家、思想家

小米董事长雷军曾说：“投身信息产业的怀抱快三十年了，我有时也在想：信息何以会具备如此强大的力量？它的力量来自哪里？我们又该如何驾驭这一力量？”作为见证中国互联网崛起的企业家，雷军依然对信息这个词有所困惑，更不要说普通人。

20 世纪 40 年代末，信息的概念开始出现，并以迅雷不及掩耳之势扩展

到很多学术领域。信息既不是宏观的，也不是微观的，它可以刻在古巴比伦人用于记录的泥板上，也可以隐藏在一段生物 DNA 中。它几乎适用于所有学科领域，这种超然的性质使得它在人类的知识体系中具有非常重要的地位。

1948 年 7 月，时年 32 岁的克劳德·艾尔伍德·香农（Claude Elwood Shannon）在 *The Bell System Technical Journal* 上发表了论文 *A Mathematical Theory of Communication*（通信的数学理论）。这篇划时代论文的发表，标志着现代信息论研究的开端，截至 2017 年 6 月，这篇文章的 Google Scholar 引用量已经超过六万。在这篇论文里，香农带给人们一个自创的新词“比特”。香农对这个新词的解释是“测量信息的单位”，到现在，比特已经和千克、分、摄氏度等一样，成为人们日常生活中常见的量纲。在香农提出的信息理论中，信息熵是一个重要的概念，可以用来定量反映信息的不确定程度。例如，在自然语言处理中，中文的静态平均信息熵是 9.65 比特，而英文是 4.03 比特，中文的信息熵大于英文，说明中文的复杂程度更高，词义更丰富，但处理起来也更困难。

在信息理论提出后，为了让人们更好地理解这一理论，香农指出需要把信息和含义区分开，不能混淆在一起。这样的考虑是出于工程和哲学两个方面的原因。从工程技术的角度来讲，香农当时致力于制造一种能够使信息得到传递的机器，该机器无须了解所传递信息的真实意义。从哲学的角度来讲，“含义”与“信息”这两个词，其实是完全不同的概念。

首先，信息本身是无意义的。麻省理工学院的知名物理学家塞萨尔·伊达尔戈在其所著的《增长的本质》一书中肯定了“信息是无意义的”这个概念。伊达尔戈认为，对人类来说，要将信息和含义区分开是很困难的，因为人们会不自觉地将自己理解的含义注入信息，并认为这是理所当然的。例如对于“1111”这一串数字，喜欢网络购物的人想到的是“双十一”购物节；单身狗们

想到的可能是光棍节；而对于学计算机的人，可能想到的是二级制的数字15。从这个例子可以看到，对于信息“1111”来说，不论是购物节还是光棍节，都是人们在无意中强加上去的含义，这些意义并不是该信息本身的一部分，但人们无法控制这种解读信息的本能。有时候，信息和含义可能会完美地结合在一起，但这也并不意味着信息本身是有意义的。

其实，信息是没有实体的，但信息又是物质性的，它能够被物理性的表达呈现出来。信息可以理解为一种物理秩序，或是物理事物的组合方式。对于汉字来说，每一个汉字都可以看作笔画的组合，或者是笔画的排列秩序，人们对其中一些组合赋予或强加了意义，才形成了文字；英文单词也是一样，是英文字母的排列秩序。显然，这样的笔画或字母的秩序远远大于汉字或英文单词的数量，那些没有被赋予含义的组合仍然是不具备意义的信息。从这个角度出发，上面所讲的信息“1111”的例子就很好理解了：1111的不同含义其实是代表了信息“1111”的某一种特定秩序。

在赛博新经济时代，信息的重要作用日益凸显，并在社会经济增长方面迸发出巨大的能量。《增长的本质》一书指出：“经济是通过实体化信息，增强人类对于知识的实际应用的系统”“经济增长的本质，是因为信息(量)的增长”。这里，信息增长促进赛博经济增长，包含了两层含义：首先，信息能够在赛博世界中快速增长。赛博时代具有有别于以往任何时代(如农业时代、蒸汽机时代)的特征，这些特征包括小世界、信息级联、幂律和长尾等，这是赛博时代所独有的由赛博带来的特征[①]。在这些特征的作用下，赛博时代中信息的增长速度超过以往任何时候。赛博世界为信息增长提供了快速路径，并且这样的路径只存在于赛博世界中。其次，促进经济增长的信息是“有序”

① 更多关于赛博特征的内容，可参见《赛博新经济：“互联网＋”的新经济时代》(清华大学出版社)。

的，即这些信息中包含了某种隐形的秩序。赛博时代中如果只有信息的快速增长是不够的，正如前面所讲，增长的信息可能是无序的，没有任何意义。物质的排列秩序是多种多样的，这些排列秩序里，有些是有意义的，有些则可能什么都不是。例如一团杂乱无章的纤维是纤维这种物质的一种排列秩序，而在另一种排列秩序下，这团纤维可以编织为毯子。显然，纤维的两种排列秩序都是信息，相对于杂乱无章的秩序，第二种排列比第一种更为“有序”，这种“有序”使信息具有了某种含义和价值，能够为人所用，这就是一种隐形的秩序。我们可以把这种信息称为“知识”。可见，这种包含在信息里的隐形的秩序才是信息的价值所在，才是这个世界所需要的，才能促进社会经济增长。

在赛博世界中，每时每刻都有大量新的信息产生，也有老旧信息消亡。对于人类社会经济系统来说，那些包含有秩序的信息（也就是知识）才是需要的，正是不断增长的知识才促进了社会、经济不断向前发展。赛博（互联网）的存在，只是信息能够快速增长的基础，这些混乱的信息需要一种力量去有序化，这股力量就是在赛博世界里不断渗透和不断演进的算法。从赛博的本质，到赛博系统中的各种上层应用，其背后都有算法的存在。算法保证了信息的有序增长。

此外，算法还在建立另外一种秩序——赛博新经济的运行规律。从经济系统底层信息、数据的产生、增长，到上层的具体应用实践，算法都在其中起着决定性的作用。信息的有序增长促进经济增长，算法决定着信息增长的秩序，同样也决定了赛博新经济系统的秩序。

新经济还需要另一只“看不见的手”

他这样做只是被一只看不见的手引导着，去促进一个并不是出自他本心的目的。

——亚当·斯密

随着互联网和移动互联网的发展，建立在其上的应用和服务，无论称为“互联网＋”还是“＋互联网”，我们都能看到它对于人类社会的巨大改变。三十年前人们很难想象，三十年后人们只需要简单地点击几下手机，足不出户就可以处理大多数日常事务。今天，一个普通人的一天很可能如此度过：

早晨，当你一觉醒来，会使用手机控制灯光亮起、享受智能厨具做好的美味早餐，浏览今日头条推送的新闻事件和好友的微信留言。出门上班，通过在线查看实时交通状况并规划出行路径和方式，可以选择搭乘网约专车，或是公共交通。在公司，按照日程安排处理公务、收发邮件、参加视频会议。到了中午，可以一边享受网上预订的外卖，一边追一追感兴趣的美剧，或是听听在线音乐。午休时间，可以打理下网络理财产品，或是逛逛电商的网站。下班后，可以尽情吃喝玩乐，在朋友圈秀自拍。

我们对上面描述的这一切已经习以为常，并将这一切理所当然归功于互联网。然而，“互联网”是一个浅层次的答案，不够准确，也不够本质。两百多年前，英国经济学家亚当·斯密（Adam Smith）在《国富论》中用“看不见的手”来描述市场机制在经济运行中的作用。与此相似，我们发现在赛博世界，在经济运行以及这些日常活动的背后，也存在着另一只看不见的手，在有条不紊地操纵着一切，维持着系统的正常运转。

而这只看不见的手，就是“算法”。日常行为的背后所涉及的一些算法如表 0.1 所示。

表 0.1　日常活动背后的算法

活动/应用	涉及的算法
手机控制智能家居	无线通信与加解密
浏览推送新闻	推荐
视频会议	包调度、音视频编解码
网络约车	匹配、路径规划、动态定价
网络购物	推荐、区块链
在线音乐、影视平台	推荐、音视频编解码
拍照美化	图像处理
网络金融	分配、区块链、匹配
交通状况查询	大数据处理

可以说，几乎所有的经济运行流程，或是人们的活动，都是由算法在幕后重新建立一种秩序。就拿在电商网站购物来说，当你选择了购买某种商品，算法会判断这种商品剩余数量是否能满足你的购买量，如果可以满足，则会将你选购的商品放入购物车；如果剩余数量不足的话，会返回一个提示信息。在结算时，算法会根据商品单价和购买数量，计算你需要支付的金额，在这个过程中，算法也会自动查询你是不是有可用的优惠券或满足打折条件。如果你选择在线支付，算法会显示相应的支付界面。付款后，算法会计算出送货时间，并显示在你的手机或计算机屏幕上。一个简化的网络购物流程如图 0.1 所示。

通过上述购物流程，可以体会到算法在人们日常活动中所起的作用，必须按照算法规定的流程来操作，否则就不会成功。其实不只是人们的日常事务，就连赛博新经济的基础——互联网，保证其正常运行的核心因素也是算法。如果您是计算机网络领域的从业者，也许对此并不陌生。这类算法有个专门的名字，称为“网络协议”。网络协议就是为了便于计算机网络中的不同

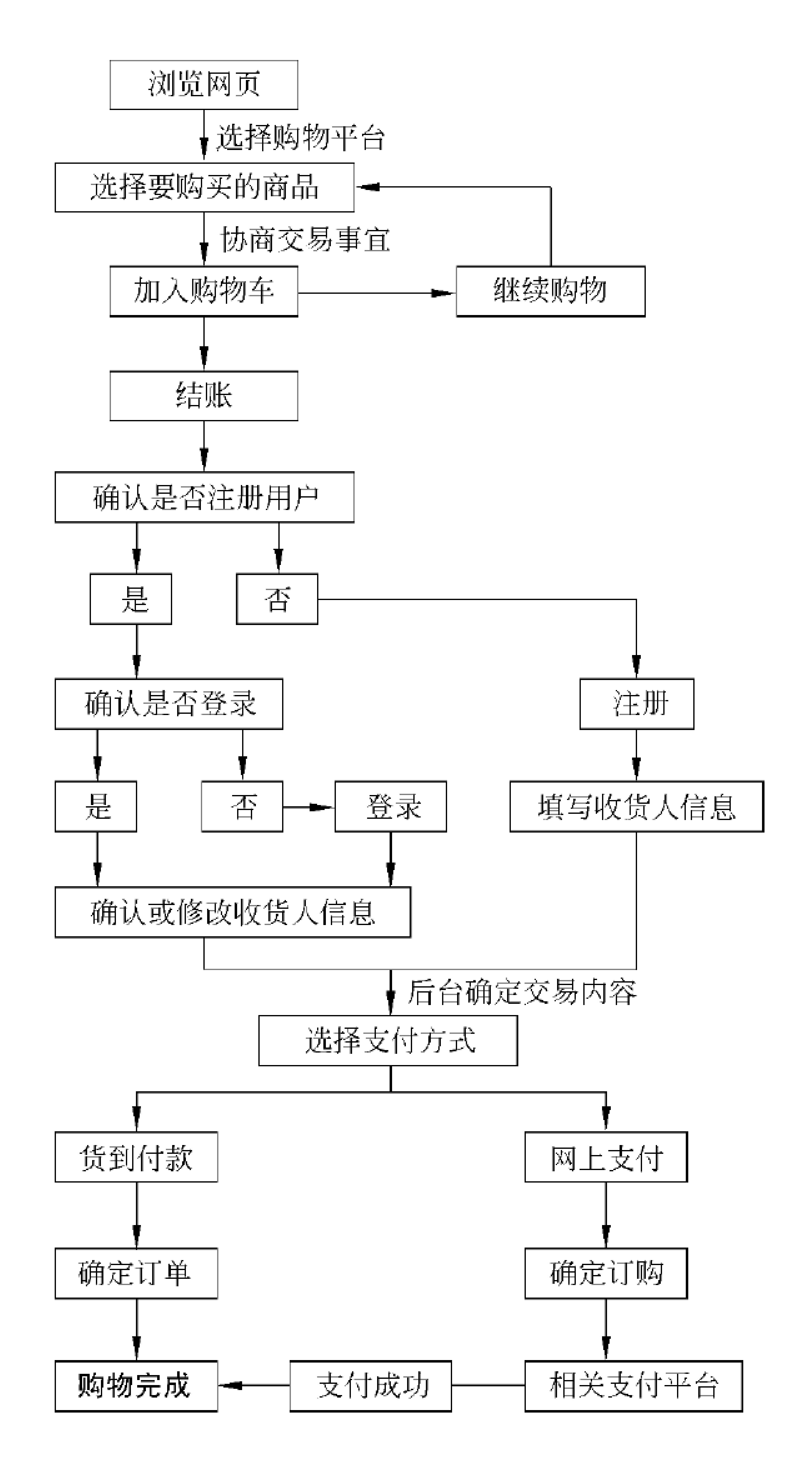


图 0.1　简化的网络购物流程图

计算机之间进行通信和数据交换而建立的所有计算机都要共同遵守的标准或约定。众所周知,互联网又称为网际网,是由各种不同的网络构成的,网络协议则是这些网络之间能够连通的保证。举个简单的例子,一位中国企业家与一位德国老板开会商谈合作事宜,如果中国企业家只懂中文,德国老板只懂德文,那他们没有办法正常交流;如果两人的秘书都懂英文,那么一方的秘

书可以先将自己上司要表达的意思转化为英文传递给对方秘书，对方秘书接收到对方信息后，再将英文转化为自己老板的母语，这样，交流就可以愉快地进行下去。在这个过程中，秘书和英语就起到了类似网络协议的作用。

一个标准的网络协议至少包括语法、语义和时序三个要素：

- 语法：规定了用户数据和控制信息的格式，包括数据出现的顺序；
- 语义：规定了各种控制信息的意义，说明通信双方应该怎么做；
- 时序：规定了事件的顺序，例如何时通信、先做什么、后做什么、传输速度等。

用上面中国企业家和德国老板开会的例子来类比的话，语法就是大家都能理解的英语语法，例如定语从句、状语从句等等；语义就是使用的英语单词和语句的意义；时序就是两位秘书事先商量好：谁先说、谁后说，先讨论什么内容、后讨论什么内容，语速是快还是慢等等。

网络协议有很多，其中最著名的莫过于 TCP/IP 协议。TCP/IP 是 Transmission Control Protocol/Internet Protocol 的简写，中文名称是传输控制协议/互联网协议。事实上，TCP/IP 协议是一个协议族，包括很多协议，例如 UDP 协议、ICMP 协议等等，TCP 协议和 IP 协议是其中最广为人知的两个协议。顾名思义，TCP 协议是为数据传输提供服务的协议，它通过一种称为“三次握手”的机制在终端之间建立连接，提供可靠的传输服务。“三次握手”机制得名的原因，是传输数据的终端之间要通过三次交互过程才能完成可靠连接的建立。有一个名为“我给你讲一个 TCP 的笑话吧”的段子生动地描绘这一过程。

路人甲：好想听一个笑话。

路人乙：你好，你想听 TCP 的笑话吗？

路人甲：嗯，我想听一个 TCP 的笑话。

路人乙：好的，我会给你讲一个 TCP 的笑话。

路人甲：好的，我会听一个 TCP 的笑话。

路人乙：你准备好听一个 TCP 的笑话吗？

路人甲：嗯，我准备好听一个 TCP 的笑话。

路人乙：Ok，那我要发 TCP 笑话了。大概有 10 秒，20 个字。

路人甲：嗯，我准备收你那个 10 秒钟长，20 个字的笑话了。

路人乙：抱歉，你的连接超时了。你好，你想听 TCP 的笑话吗？

好吧，这个笑话有点冷，我们需要知道的是网络中所有的“行为”，例如建立连接、发送/接收数据等等，都是由标准化组织制定的各种网络协议来规范的。这些网络协议不能直接搬到网络上用，还需要由码农们用机器语言把这些协议“翻译”成各种网络终端能够“看懂”的形式，也就是大家俗称的“程序”。在这些程序里，包含了各种各样不同功能的算法，正是这些算法保证了网络协议中所描述的功能能够实现，例如路由协议的实现依赖于最短路径算法，传输控制协议的实现依赖于排队算法等等。一般来说，网络协议的实现需要多个算法共同支持。有时候，网络协议本身就是一个算法，例如三次握手协议。很多时候，网络协议之间还存在复杂的依赖关系，这里就不多讲了。

是的，我们使用的每一个计算机软件、程序、APP，背后都是算法。也许我们常在新闻报道里见到“程序”“应用”“软件”“算法”等意义相近的词汇，这在媒体报道里很少加以区分。

不过，为了保证其严谨和准确性，我们在本书会使用算法，而不是“程序”或是“软件”。算法、程序、软件，是三个既有区别又有联系的词汇。算法是针对某一问题的解决方案的准确描述。例如计算机领域的排序算法描述的是，给定一列数，按照由大到小或由小到大的顺序输出：如果输入是(5，3，9，2，6)，则输出是(2，3，5，6，9)或(9，6，5，3，2)。算法可以使用自然语言来描述，当然也可以使用计算机语言或是数学语言。算法一定要在有限的步骤中得到问题的结果，即满足有穷性。而程序就不一样了，程序不一定满足有穷性，

它可以进入无限循环的状态,例如计算机操作系统,只要不关机或者系统不遭到破坏,它就会一直运转,等待新的任务到达。此外,程序是使用编程语言写成的,用于实现某种特定目的的一组计算机指令。如果将一个算法使用编程语言进行描述,就成为一个程序,该程序是这个算法在计算机上的特定实现。软件,则是程序的有机集合体,软件中可以只有一个程序,也可以是多个程序的集合,程序从属于软件。例如手机上或计算机里最简单的计算器,就只包含数值计算程序;而 Office 办公软件则包含了 Word、PowerPoint、Excel 等多个程序。对比算法、程序、软件这三个词,可以看出算法的含义是最明确的,程序或软件只是算法的某种特定实现或这些实现的集合,而且程序或软件包含的范围更大,不如算法准确。算法描述更准确、更本质。

被算法唤醒的新经济

生产率不等于一切,但长期看它几乎意味着一切。

——保罗·克鲁格曼,经济学家,诺贝尔经济学奖获得者

从经济系统的角度看,赛博经济系统继承了传统经济系统的各个组成部分,并赋予了它们新的特点,形成了赛博下的生产系统、交换系统、消费系统和金融系统,如图 0.2 所示。

从上至下贯穿赛博经济系统的算法包括八类,分别是推荐算法、分配算法、匹配算法、动态定价算法、区块链技术及相关算法、大数据处理算法、数据交易算法、隐私保护算法。同时,每一类都包括针对不同具体问题的算法,以及数量不菲的侧重不同方面的变种。

由于经济系统不是孤立运行的,它必然会造成同一类算法可能既存于这

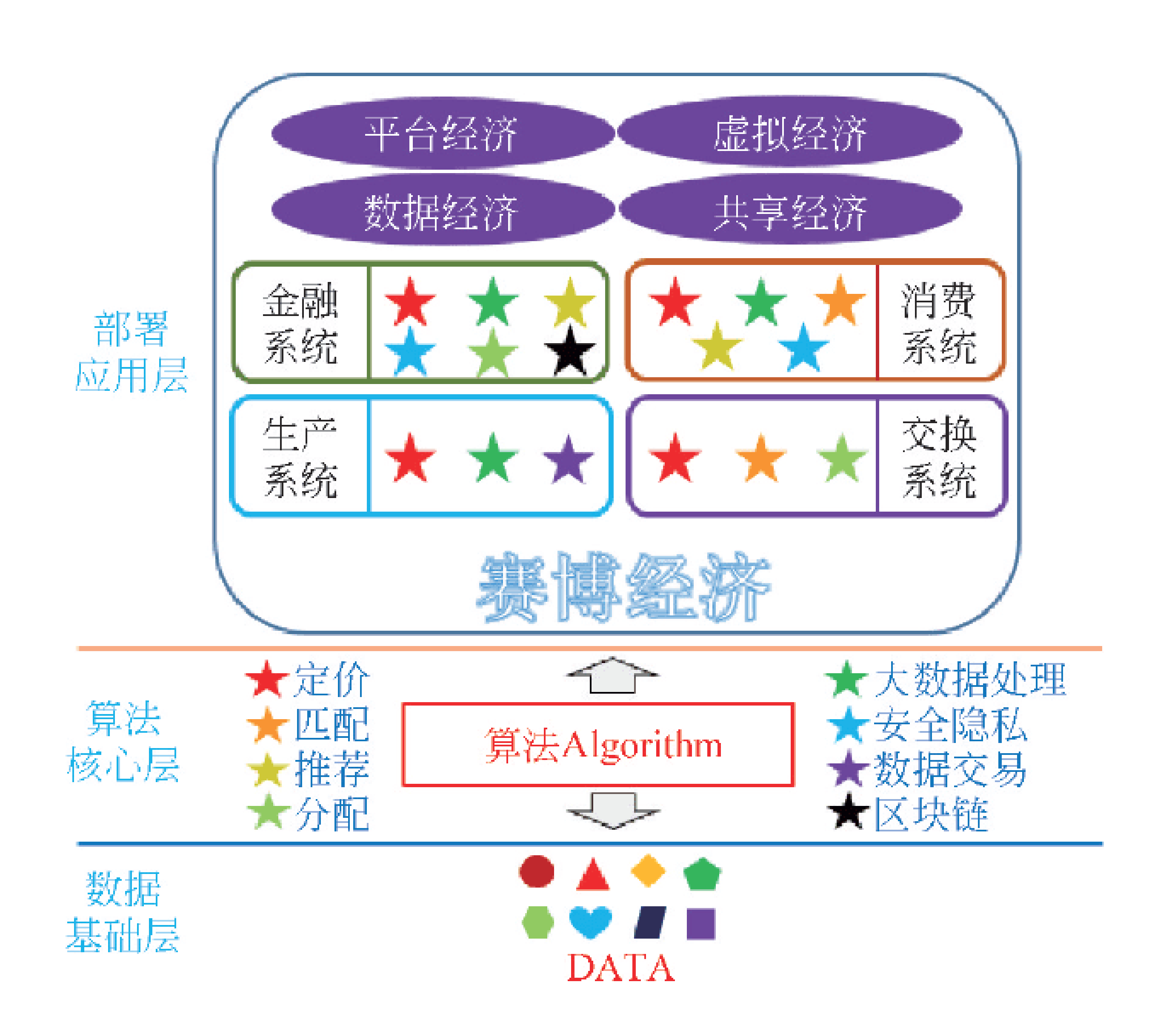


图 0.2　算法定义经济总体结构图

个子系统，又跟那个子系统有联系。我们尝试把这八类算法按支撑关系对应到四个经济子系统中，会看到在金融系统和消费系统中包含的算法多于另外两个系统。这是因为，金融系统和消费系统从来就是经济体系中最活跃、最有生命力和创造力的两个系统，与实际生活联系最紧密，相应地，衍生的经济活动和关系最复杂，需要更多的算法提供支撑。不过这并不意味着支撑生产系统和交换系统的算法就少，要知道种类相对少并不代表数量就一定少。

以消费系统为例，动态定价算法大量应用于日常的网络约车、团购等应用。人们在淘宝、京东、亚马逊等众多电商的购物数据，特别是“双十一”等节日消费，构成了人们的消费大数据，电商要从这些繁杂的数据中推断用户的个人喜好、群体趋势，必然要用到大数据处理算法。在赛博时代，人们的很多行为都搬到了互联网络上，例如征婚交友、股票买卖，这些都是匹配算法的用武之地。推荐算法就更不用说，是广告商和电商最关注的算法之一，因为推荐精准度往往决定了销售量的高低。无论是以前，还是现在，安全和隐私问题一直受到人们很大的关注，在赛博时代更是上升到一个前所未有的高度，

交易安全、隐私保护等算法可谓任重道远。

赛博经济时代，受到赛博世界独有的、不同于以往任何时代的“小世界”、幂律和长尾、无所不能的计算等特点的影响，生产活动变成以消费者的需求为导向，满足消费者的不同个性化需求为目的，从而展现出新的经济形态特征，比如数据经济、共享经济、平台经济和虚拟经济。显然，这四个方面的特征跟赛博经济系统的特点密不可分，仍然离不开算法的支撑。

从数据基础来看，在赛博时代下，每天产生的数据用海量已不足以形容。据报道，搜索引擎谷歌(Google)现在每天要处理来自全球的超过 55 亿次的搜索请求；早在 2012 年 1 月，视频网站 YouTube 每天为观众播放影片的次数就达到惊人的 40 亿次，服务器每分钟要应付长约 60 小时的影片上传量，也就是说在 1 秒钟的时间里要处理长度为 1 小时的影片。网络公司思科(Cisco)预测，全球互联网总流量将在 2016 年超过 1ZB，而这个数值在 2019 年还要再翻一番。

ZB 是容量单位，全称是 Zetta Byte，它的大小关系是 1ZB＝1024EB＝1024×1024PB＝1024×1024×1024TB＝1024×1024×1024×1024GB，这个数据量有多大呢？按全球人口 70 亿计算，人均产生约 157GB，也就是约 160 843MB 的数据，这相当于每人拍了 32 168 张数码照片(每张照片按 5MB 计算)，或是拍摄 7 小时长的 4k/30fps 高清视频(按 1 分钟 4k/30fps 视频约 375MB 计算)。

互联网时代以前，数据量还不大，用人工的方法大致上还能应付过来。但今天我们必须面对这些多到无法想象的数据，其中还包括很多无秩序的数据，靠人工显然是无法处理的，要想发现或挖掘出数据中蕴含的价值，只能求助于大数据处理算法。也就是说，没有算法，数据累积得再多也只是一堆数据而已，没有什么用处；只有算法出现了，数据中隐藏的有用信息(秩序)才能被发现，数据也才能变成“金矿”。

在前面的介绍中我们已经了解到，从信息的角度出发，赛博经济是信息增长更快的经济。到这里，我们已经明白，信息有序增长与快速增长的核心原因是算法。因此，从算法的角度来定义赛博新经济可能更为准确。

算法贯穿了赛博经济系统的方方面面，支撑整个系统运转，核心地位不言而喻。如果说，互联网的出现与发展，形成的是赛博经济系统的"身体"，那么，算法形成的则是赛博经济系统的"灵魂"，这两者相辅相成，缺一不可，构成了一个有机整体。在此基础上，我们可以在赛博新经济概念的基础上再前进一步，提出算法定义经济（Algorithm Defined Economic System，ADES）的概念，即：

算法定义经济是指以算法为核心的、以信息（包括知识和数据）为资源、以网络为基础平台的一种经济形态，在其中，算法决定了信息增长的秩序，同时它贯穿了经济系统的所有组成部分和流程，支撑并控制系统中各种经济活动以及所形成的各种经济关系，决定了经济系统的秩序。

我们从算法定义经济的概念里，也可以看出这种经济类型的一些特性：首先，它是一种知识型、科技型经济。算法本身是一种完成某种特定功能的、高度凝练的知识，属于科学技术的范畴，自然而然地就给经济系统也打上了科技的烙印。其次，它是一种高渗透型的经济。从底层的基础平台、数据的产生和处理，到经济系统的组成、运行的全过程，算法的身影无处不在，这是一种灵魂与躯体的关系，高渗透是一种天生的属性。同时，它是一种"软"性经济。历史上曾经出现过的农业经济、工业经济等经济形态，农业、工业都是具体的、实在的概念，是"硬"属性。而算法从本质上讲是一种知识，属性偏"软"，不过，这种"软"经济却具有相当的硬实力。在这种"软"性经济形态下，经济系统的各组成部分都得到极大的强化和优化，经济活力和增长速度都超过以往任何一种经济形态，充分诠释了"科技是第一生产力"这句名言。

由于算法自身所具有的独特属性,使得算法定义经济具有高度自动化、高可扩展、高鲁棒性等特征。第一,它是一种高度自动化的经济形态。从网络基础,到上层各类应用过程,都由算法掌控,减少了人为因素的干扰,系统运行效率高。第二,它是一种高可扩展的经济形态,经济活动和行为由算法定义和控制,意味着经济活动的增加和消亡,从算法的角度来讲非常易于实现。第三,它是一种高鲁棒性的自适应的经济形态,算法是随信息技术的前进而动态发展的,在这个过程中可以不断自身修复和迭代,提升效率和增强健壮性,从而带动整个系统健康发展。在当前,算法定义经济的一个典型例子就是赛博新经济。

正在崛起的算法帝国

进化可能会在不同系统中创造出用来解决相同问题的不同算法。

——黛博拉·戈登,生物学家

讲算法,就必然要说到人工智能,也要讲到大数据,很难把这三者割裂开。在实际中,很多人经常将算法与人工智能混淆在一起。我们认为,算法是基础,人工智能是表现,人工智能是建立在算法高度发达的基础之上。同样,大数据之所以成为“金矿”,也是因为有能够处理、挖掘大数据的算法出现。

算法的发展历史远远长于互联网的历史,早在古巴比伦、古埃及时代就已经有了算法的记载。从最初的萌芽到现在,算法从来没有停止它的步伐,一直在进化发展。在早期,算法可能更多地聚焦在数学问题的计算及其相关应用上。从 20 世纪三四十年代开始,得益于现代计算机和互联网的发展,算

法的发展有了长足的进步,加快了向其他领域渗透的步伐;算法的规模也不再局限于在单台终端,大型分布式算法纷纷出现。特别是 20 世纪末,互联网浪潮汹涌而来,再加上后来移动互联网的爆发式发展,互联网已经成功“入侵”人类社会经济生活的方方面面。可以说,如果没有互联网这个新丁(相对于算法来说)的出现,算法也无法达到现在“一统天下”的高度。总的来说,互联网从两个方面加速了算法成就其统治地位。首先,互联网填补了亘古以来就存在的不同地域间巨大的时间和空间沟壑,将世界万物连在一起,世界被推平了。世界变平以后,时空距离大大减小,人类经济社会联系更加紧密,时效性大大提高,这使得经济系统的运行不断加速。在这样的情况下,原来的经济系统不断被算法改造,算法深度嵌入经济系统的运行。第二,互联网使得产生信息(数据)与收集信息(数据)的效率同时提升,不仅信息(数据)产生地更快、更多,信息(数据)的收集也变得越来越快。这种情况的出现,从另一个方面促进算法的进步,能够处理海量数据的算法不断出现,也使得人们越来越依赖大数据处理的结果。古时的“兵马未动,粮草先行”,在赛博时代变成了“兵马未动,数据先行”。

最近,现实生活中也出现了越来越多的算法成为主导的例子。据说,2016 年美国总统大选,特朗普如愿登上总统宝座,就是借助了算法分析的结果。2016 年 9 月,特朗普的竞争团队给英国的一家数据分析公司 Cambridge Analytica 支付了 500 万美元(也有人说是 100 万美元)的咨询费用。这家公司使用先进的大数据处理算法帮助特朗普分析美国选民的行为。根据分析公司的分析结果,特朗普的团队能够精准把握选民的心理,投放竞选广告时也能做到有的放矢,在不知不觉中占得先机,这 500 万美元花得很值。而反观希拉里的团队,还是采用传统的民意调查数据,同时也过度依赖所谓专家的意见,最终只能黯然咽下失败的苦果。对比特朗普和希拉里的竞选,可以看到特朗普依靠大数据和算法,有两个优点:一是分析公司获取的大数据来

源更丰富、更客观、也更贴近真实,而民调数据数据量小且比较片面;二是算法分析更客观,不带入任何主观性,而专家分析则往往带有专家本人的倾向,引入了主观干扰。在这次大选中,比尔·盖茨、巴菲特、扎克伯格和马云等领袖人物押错了宝,可谓追悔莫及。

在赛博经济时代,算法已经逐渐“上位”的事实也可以从“全球最大市值的上市公司”的排名变化中窥见一二。据有关报道,在 2001 年,全球最大市值上市公司的 Top5 分别是通用电气、微软、埃克森美孚(石油)、花旗银行、沃尔玛(零售)。到了 2006 年,这个排名变为埃克森美孚(石油)、通用电气、道达尔(石油)、微软、花旗银行,前五名中,科技公司的代表仍然是微软。再过 5 年,到了 2011 年,Top5 的排名变化为埃克森美孚(石油)、苹果、中石油、壳牌(石油)、中国银行。这个时期,科技公司的代表仍然只有一家,只是从微软变成了苹果。不过,可喜的是,中国企业中石油和中国银行上榜!时间来到 2016 年,Top5 排名发生了翻天覆地的变化,排名前 5 的是苹果、谷歌、微软、亚马逊、Facebook。可以看到,这 5 家公司都是清一色的互联网科技公司,算法是这类公司竞争力的集中体现。而传统的钢铁、石油、银行、地产等行业的跨国企业则悄然走下了神坛。

在新经济时代,得益于算法的发展和支撑作用,应用层面的各种新玩法和新模式层出不穷,经济系统展现出勃勃生机。不仅如此,底层基础网络平台也不甘寂寞。随着算法的发展,人们对未来网络的架构又提出了很多新的设想,SDN(Software Defined Network,软件定义网络)就是其中有代表性的一种。SDN 的设想是将网络的控制层面和数据层面分开,由用户根据实际需要,通过软件(实际上是算法)来控制和管理底层的网络设备,完成数据流的路由、转发等功能。这一设计赋予了互联网极大的灵活性和敏捷性,能够契合赛博时代下瞬息万变的业务环境需求。SDN 一经提出就得到了各国电信运营商和互联网设备生产商的高度关注和支持。可以看到,SDN 其实也

是互联网算法发展到相当高度的产物，如果 SDN 完全实现，将进一步巩固算法在系统中的核心地位。

算法不会停止它前进的脚步。在算法的驱动下，将来必然会出现更高级、更有活力的经济形态，不断将社会经济推向更高的水平。在不远的将来，我们会越来越深刻地感受到，算法崛起，势不可挡。

目　录

第 1 章　算法到底是什么

人类其实一直就生活在算法的世界里。关于算法的智慧早就充斥在中国的万里长城、古巴比伦的空中花园、阿尔忒弥斯神庙和埃及金字塔里，充斥在“曹冲称象”“田忌赛马”等典故里，也充斥在今天时刻运转着的互联网、奔跑的汽车以及天空中的各种飞行器里。大约是从十几年前起，人们开始频繁提到“算法”这个词，然而，那时大多数人可能还不了解算法的准确含义，或者认为，这是属于搞数学或者计算机科学的专家才应该明白的事情。到了今天，算法已经成了人们耳熟能详的一个词语，社会生活的各个角落和我们生命的每分每秒都与算法紧密相连。算法不仅存在于人们的手机和笔记本电脑中，还存在于居住的房屋、使用的电器、乘坐的汽车火车，以及小朋友爱不释手的各种电子玩具中。现代的算法能够安排人们每天的日程，能够处理复杂的金融交易，还能管理和经营企业，为人们提供各种娱乐和生活便利。我们已经不敢想象，如果突然有一天，所有的算法都停止运转，人类的文明秩序会不会立刻崩塌？

如何定义算法

今天大部分的软件都很像上百万块砖堆叠在一起组成的埃及金字塔，缺乏结构完整性，只能靠强力和成千上万的奴隶完成。

——艾伦·凯，图灵奖获得者

古时的计算方法

雄伟的非洲大陆上，世界第一长河——尼罗河自南向北流贯非洲东北部，它起源于大湖地区的卡盖拉河，最后注入地中海。尼罗河有定期泛滥的特点，每年 8 月到达最高水位，洪水来时会淹没两岸的农田，洪水退去留下厚厚的淤泥，形成肥沃的土壤。日积月累，形成了现在的尼罗河三角洲，这里物产丰富、人口密集，孕育了古埃及文明。

在尼罗河下游岸边，散布着几十座大大小小的金字塔。金字塔是古埃及劳动人民高超建筑技艺的杰作，大约在公元前 3000 年左右开始出现，是世界八大建筑奇迹之一。著名的胡夫金字塔，建成时高 146.5 米，底座每边长 230 多米，三角面斜度 52°，塔底面积 52 900 平方米；它的塔身由 230 万块石头块砌成，这些石块平均重 2.5 吨，最大的重达 160 吨。有学者估计，如果把这些石块凿碎，铺成一条一尺宽的道路，大约可以绕地球一周。更令人惊叹的是，金字塔塔身的石块之间，没有任何水泥之类的粘着物，而是像搭积木一样由这些大小不一的石块垒起来的；这些石块各个面都磨得很平，虽然至今已过去数千年，但人们也很难把薄薄的刀刃插入石块之间的缝隙。

哈瓦斯(Zahi Hawass)是埃及著名的考古学家和金字塔专家,多年来一直致力于金字塔、古埃及遗迹及其历史研究。他的研究表明,古埃及人在建造金字塔前,很可能绘制了详细的平面设计图,这其中就涉及大量复杂的计算和测量。在那个时代,古埃及人已经开始使用绳子来丈量土地。有一个专门的工种称为职业结绳人,他们的工作就是在测量用的绳子上,根据规定的长度打出等间隔的绳结,负责修建的工人使用这种打结的绳子进行丈量。有说法称,很可能是这些职业结绳人最先发现某些长度固定的三条绳子能够构成直角三角形,这些绳子包括由 3 个、4 个、5 个等间隔的绳结长度组成的,也包括由 5 个、12 个、13 个等间隔的绳结长度组成,这正好是勾股定理描述的直角三角形三边长。

而中国最早记载勾股定理的古籍是《周髀算经》,据记载,大约是周武王灭商(公元前 1046 年)的时候,一个叫商高的人发明了勾股测量术,也就是我们平时所说的"勾三、股四、弦五",比古埃及人略晚一些。古埃及人还需要解决的另一个难题是如何准确画出直角,因为金字塔的地基必须是正方形,它的四个角必须是严格的直角,如果稍有偏差,就会导致整个建筑走形,甚至垮掉。史学家研究表明,古埃及人可能是这样做的:先在地上打进两个木桩,然后绷紧木桩间的绳子,这样就画出一条直线,成为金字塔的一条边线。然后,在两个木桩上各系上一条绳子,绳子的长度要超过两个木桩距离的一半。拉紧绳子的末端,以木桩为原点转动,画出两条相交的圆弧来。过这两条圆弧的交点,画出另一条直线,与先确定的那条直线相交,夹角就是准确的直角。这里的后一条直线,就是地基的另一条边线,如图 1.1 所示。

从古埃及人建造金字塔的过程中,我们能隐隐约约看到计算方法的萌芽。正是这些实践应用中的需求,使得数学和计算方法日益受到重视并开始发展起来。

除了古埃及,另一个古文明——古巴比伦,也留下了丰富的关于计算的

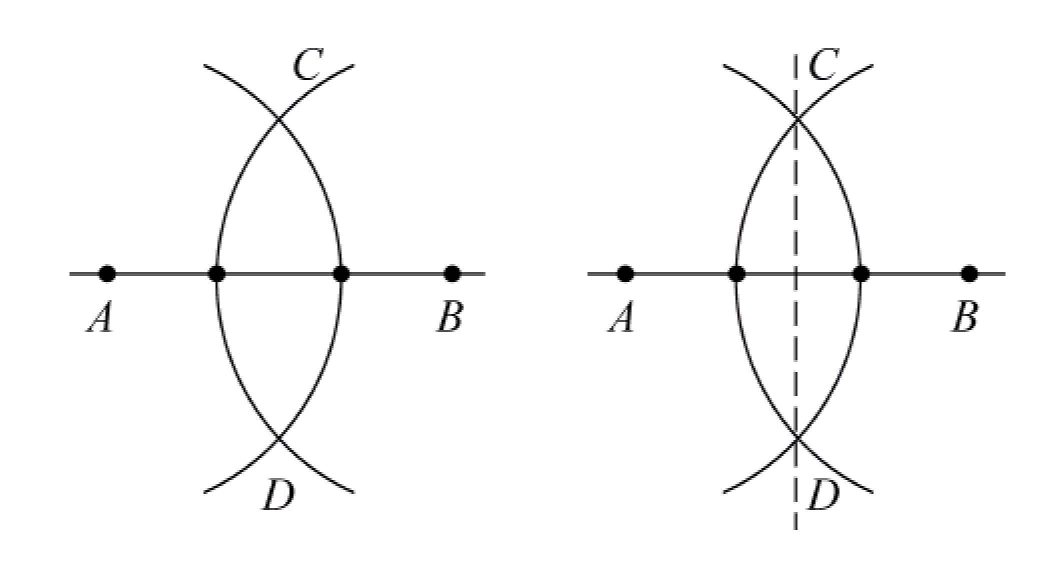


图 1.1 古埃及人画直角示意图

数学遗迹。大约公元前 18 世纪，古巴比伦王国出现在美索不达米亚平原地区，大致位置是在现在的伊拉克共和国境内。古巴比伦当时使用一种巴掌大小的泥板来记录各种信息，这些信息使用一种有棱有角的楔形文字，压印在这些泥板上。这种文字是由约公元前 3200 年的苏美尔人所发明，是世界上最早的文字之一。

通过近两百年对美索不达米亚的考古发掘，以及语言学家对大量泥板文献成功地译读，人们终于知道楔形文字是已知的世界最古老的文字。它是由古代苏美尔人发明，阿卡德人加以继承和改造的一种独特的文字体系。巴比伦和亚述人也先后继承了这份宝贵的文化遗产，并把它传播到西亚其他地方。从 17 世纪开始，探险家和考古学者就曾从两河流域一带破碎的陶器，以及石雕和泥板的残片上发现了这些奇异的文字符号。到了 19 世纪，考古学家先后发掘出数千块印有楔形文字的古巴比伦时期的泥板。经过几代人的不懈努力，大部分泥板上的楔形文字被解密。让人大吃一惊的是，有数百块泥板上都是记录的关于计算和数学方面的内容，包括乘法表、倒数表、平方和立方表等，如图 1.2 所示。

古巴比伦文明当时使用的是六十进制，不是现在常用的十进制。在很多泥板上，都能发现算法的踪影。例如，在图 1.3 所示的泥板上就记录着这样的问题：一块长方形土地的面积加上长与宽之差是 33(六十进制，按十进制

$524551=2(60^3)+25(60^2)+42(60)+31$



图 1.2　古巴比伦楔形文字和出土的泥板

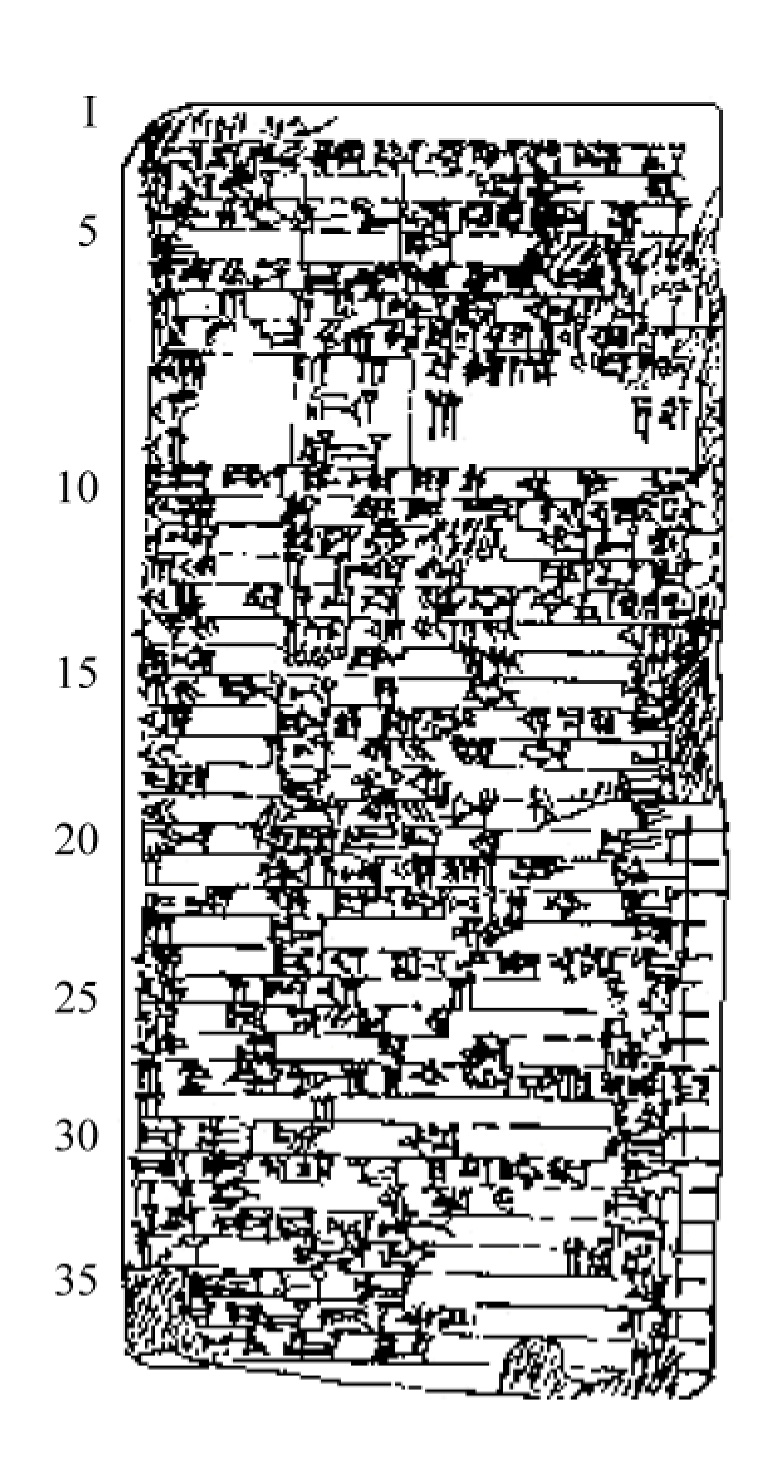
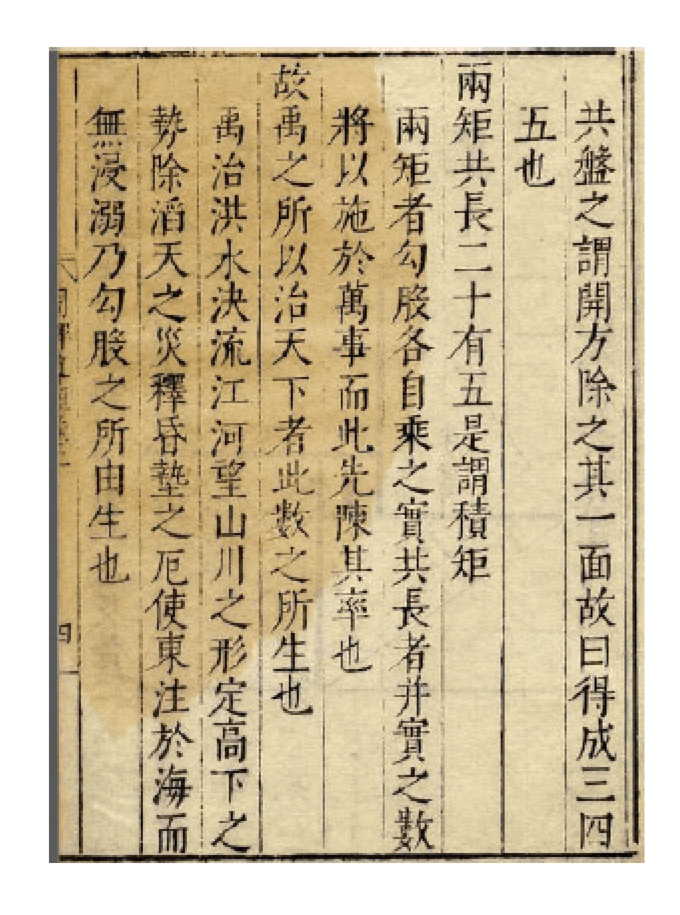
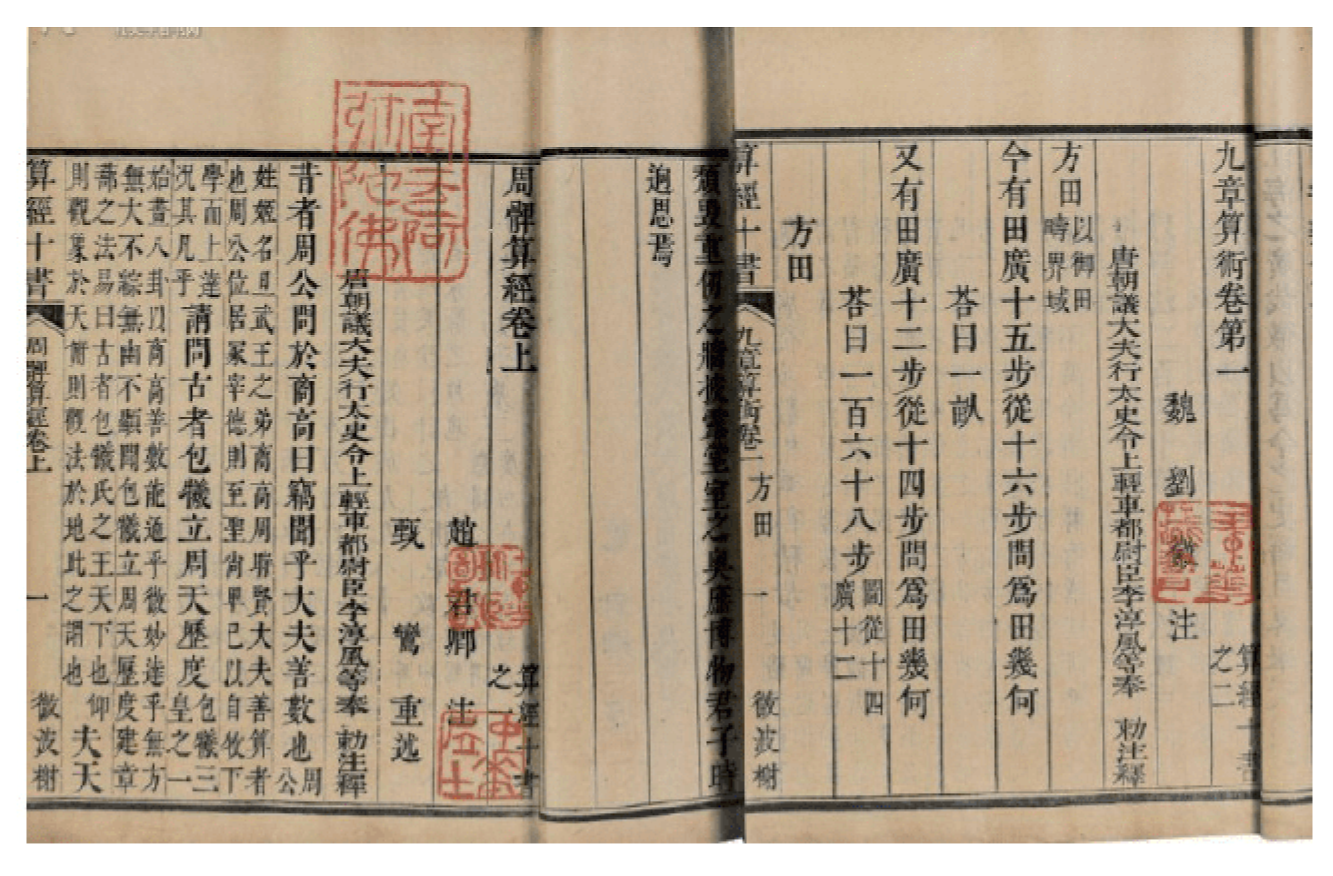


图 1.3　记录数学问题的古巴比伦泥板

就是 183),而长与宽之和为 27,问这块地的长、宽以及面积是多少?古巴比伦人的解法书写在这块泥板的第 6～29 行,他们也算出了正确的结果,这块地的长为 15,宽为 12。

再来看看我国古代数学计算方法的发展情况。我国关于算法的研究最早可以追溯到公元前 1 世纪,在《周髀算经》里就描述了很多数学计算和天文学方面的内容。在我国古代,一般涉及数学计算方面的著作,作者大都会冠以“算经”二字。在唐代,国子监设立的算学馆就将当时已有的十部著名的数学著作作为教材,统称为“算经十书”,这十部书是:《周髀算经》《九章算术》《孙子算经》《五曹算经》《夏侯阳算经》《张丘建算经》《海岛算经》《五经算术》《缀术》和《缉古算经》。在这些早期的数学著作中,记载了大量的计算方法,比如前面提到的勾股定理,还阐述了分数问题、“盈不足”问题,甚至包括负数等,如图 1.4 所示。


共盤之謂開方除之其一面故曰得成三四
五也
兩矩共長二十有五是謂積矩
兩矩者勾股各自乘之實共長者并實之數
將以施於萬事而此先陳其率也
故禹之所以治天下者此數之所生也
禹治洪水決流江河望山川之形定高下之
勢除滔天之災釋昏墊之厄使東注於海而
無浸溺乃勾股之所由生也


九章算術卷第一
魏 劉徽 注
唐朝議大夫行太史令上輕車都尉臣李淳風等奉 勅注釋
方田 以御田疇界域
今有田廣十五步從十六步問爲田幾何
荅曰一畝
又有田廣十二步從十四步問爲田幾何
荅曰一百六十八步 圖從十四廣十二
方田
算經十書 九章算術卷一 方田 一 微波榭

周髀算經卷上
趙君卿 注
甄鸞 重述
唐朝議大夫行太史令上輕車都尉臣李淳風等奉 勅注釋
昔者周公問於商高曰竊聞乎大夫善數也
請問古者包犧立周天歷度
算經十書 周髀算經卷上 一 微波榭

图 1.4 中国古代数学计算著作

有趣的是,古人不仅仅把计算方法用在数学问题上,还应用在其他领域。中国唐代有一个“杨损考吏”的故事。这个故事是说唐代有一位清正廉明的尚书官杨损,有一次需要从两名下属中选拔一人到当时的财务部任职。经过考评,两名候选人都很优秀,难分高低。主管这项工作的官员难以抉择,于是

向杨损请示。杨损听完考评介绍，想了一会儿，说道："这个部门的办事人员，需要具备比较扎实的计算能力，这样吧，咱们出一道计算题，谁先做出正确答案，就选谁。"于是，杨损给出了一道计算题："有人无意中听到几个盗贼在分赃，偷的是金砖。如果每个人分 6 块，就多出 5 块；如果每个人分 7 块，就会差 8 块，问有几个盗贼在分多少块金砖？"最后两名候选人中先算出正确答案的得到了提拔，杨损也因办事公道、任人唯贤而名声流传。

近代的算法

布伦瑞克(Brunswick)是德国中北部一个的城市。据考，布伦瑞克最早并不是城市，而是几个互不相关的居民点，后来这些居民点各自发展，最后连接到一起，才形成了后来的城市。另外一个说法则是说布伦瑞克是由来自萨克森王朝的两兄弟布鲁诺和丹克瓦尔德奠基而成的。在 1787 年的一天，布伦瑞克一个偏僻小镇上的小学里，三年级的一个班级正在上数学课。这是该校首次创立的一个班，与其他班级的不同之处是，这个班要讲授数学课，类似于现在大家耳熟能详的数学实验班，虽然孩子们在此之前根本就不知道什么是算术。上课的老师名为布特纳(Buttner)，他对孩子们的态度并不算好，对教学也不太认真。在布特纳看来，自己在这个穷乡僻壤教书真的是怀才不遇，是上帝跟他开了一个天大的玩笑。

一天，他睡眼惺忪地来到教室，头一天晚上他跟几个朋友玩牌玩了一个通宵，现在恨不得马上倒在床上美美睡上一觉。"怎么才能把这节课对付过去呢？"布特纳在暗暗盘算后，决定让这帮小子算一节课的计算题。于是，布特纳在黑板上写下了那道举世闻名的加法题："1＋2＋3＋…＋100"，并告诉同学们今天上课的内容就是算出这道题的结果，如果下课没做完就作为家庭作业。就在布特纳自以为"阴谋"得逞，准备回自己的屋子睡觉时，一个 10 岁

的小男孩站了起来说出了正确答案“5050”。布特纳大吃一惊，在他说话这短短几分钟，眼前的男孩居然算出了答案。布特纳追问男孩是怎么算的，这个男孩不慌不忙地说：“我观察到 1＋100＝101，2＋99＝101，3＋98＝101，……，50＋51＝101，一共有 50 个 101，所以答案是 50×101＝5050”。看到这里，读者会觉得这个男孩是个天才啊！不错，这个男孩就是以后鼎鼎大名的数学家高斯(C. F. Gauss)。不过也有人质疑这个情况是否属实。根据数学史学家贝尔(E. T. Bell)考证，当时布特纳给孩子们出的题没有传说中这么简单，实际上他出的是“81 297＋81 495＋81 693＋…＋100 899＝?”这是一道更复杂的等差数列求和问题(项数是 100，公差为 198)。据贝尔讲，高斯在晚年时常常向人们炫耀这件事，说布特纳刚写完试题，他就算出了答案，而其他小伙伴才刚加完前面几个数。

不管高斯当时计算的题目是简单的还是复杂的，这个故事都告诉我们，高斯从很小的时候就已经开始注意把握更本质的数学方法，也就是我们所说的“算法”。

这两年有一个非常火的真人秀节目，名为《奔跑吧！兄弟》，居然有一集里也出现了算法。在 2015 年的第二季第二集《超体元素保卫战》中，兄弟团被黑衣人关进了密室，要逃出密室需要解决各种设置的难题，其中一间密室的逃脱条件是需要解决一道名为“鸡兔同笼”的数学计算题。“鸡兔同笼”问题是中国古代著名的趣题之一，也出现在我们小学四年级的数学课本(人教版)中，相似的问题还包括和倍(差倍)问题、植树问题、鸽巢问题、行程问题，也许不少朋友都曾在小学期间被这些问题“虐”得很惨。《孙子算经》也记载了这个有趣的问题：“今有雉兔同笼，上有三十五头，下有九十四足，问雉兔各几何?”翻译成现在的白话文就是：“鸡和兔关在同一个笼子里，它们一共有 35 个头，94 只脚，问鸡、兔各有多少只?”对这个问题，相信很多人的第一反应就是使用列方程等传统解法。这里，我们来看一个思路非常奇葩酷炫的

“抬脚法”，感受一下精妙算法的魅力。“抬脚法”的思路是这样：假定笼中的鸡和兔都是训练有素，指挥鸡和兔同时抬起一只脚，一共是 35 只脚；再指挥它们同时抬起第二只脚，这又是 35 只脚，因为鸡只有两只脚，所以鸡就只能就坐地上了，只有兔子还用两只脚站着；此时抬起来的脚一共 70 只，剩下的 94－70＝24 只脚都是兔子的，所以兔子有 24÷2＝12 只，鸡有 35－12＝23 只。

算法因计算数学而起，而现代算法的应用范畴早已远远超出了数学计算的范围，已经与每个人的生活息息相关。今天我们看到大部分轿车都配备了电子驻车制动系统 EPB(Electrical Park Brake)，给车主带来了更安全简捷的驾驶体验。但是，大多数人可能并不清楚，其实 EPB 系统也是由算法来控制的。

EPB 是一种用电子控制的方式来实现停车制动的技术，它将平时行车过程中的临时制动与停车后长时间制动功能整合在了一起。简单地说，这套系统的工作原理与机械式手刹相同，都是通过在刹车盘与刹车片之间产生摩擦力来达到控制停车制动的效果，只不过控制方式从之前的机械式手刹拉杆变成了现在电子按钮。

自动挡汽车临时制动的过程一般是这样的。开车到路口遇到红灯时，在没有配备 EPB 系统的汽车上，驾驶员一般是踩下刹车，待车辆停稳后，拉起手刹，将挡位拨到空挡，然后松开脚刹；等绿灯亮起时，换到前进挡，踩油门驶出。当然，嫌麻烦的驾驶员也可以选择一直踩着脚刹不放，只是稍微累一点。配备有 EPB 系统的汽车就不一样了。同样是在路口遇到红灯，驾驶员踩下刹车，车辆停稳后，电子手刹自动启用，驾驶员松开脚刹即可；等绿灯亮起，直接踩下油门，车辆向前驶出。EPB 系统确实让驾驶变得简单了许多，它背后的算法是这样工作的：当驾驶员踩下刹车时，通过传感器感知速度的变化，车载电脑中的算法就知道车辆现在正在减速，当车辆完全停下一段时间(比

如2秒钟），驾驶员仍踩着刹车，算法就认为需要停车了，然后自动启用电子手刹，“告诉”电机卡紧刹车片来达到制动的目的，同时，算法让变速箱与发动机暂时分离，切断发动机向变速器输入的动力，因此驾驶员可以松开脚刹。此时即使在坡道，电子手刹也能控制车辆，避免不必要的滑行。当启动时，驾驶员轻轻踩下油门踏板，根据传感器传来的信号，算法判断要前进了，会“通知”变速箱与发动机逐渐结合，车辆产生向前的驱动力，算法通过各种传感器提供的信息计算，当驱动力大于行驶阻力时就会自动释放刹车片，从而使汽车能够平稳起步。此外，通过设置EPB，还可以让汽车在较长时间停止时（如道路堵塞），自动关闭发动机，在踩下油门踏板时自动点火发动，达到节约燃油的目的。这个过程看似复杂，实际上背后的算法在一瞬间就完成了，一切都发生得那么自然，似乎它本来就应该是这个样子。

什么是算法

其实，人们在平日生活中无时无刻不在使用算法，只是人们早已习以为常，以至于并没有注意到。在计算机科学家看来，人类的思维其实就是算法，思维决策的过程其实就是算法的运行过程。试想一下，每天早晨起来，决定穿职业装还是休闲装、穿靴子还是运动鞋、拎哪个颜色和品牌的包包等等，其实都是大脑思维根据今天的天气、温度、需要处理的事项等参考量，通过各种分析处理得出的一个输出结果；再比如，早餐吃什么、出门是自己开车还是叫出租车、走哪条线路等等，都是相似的算法过程。只不过这个过程与生俱来，没有人会特别关注。

在中国古代，算法被称为“术”，比如前面提到的《九章算术》，或者三国时期魏国的数学家刘徽提出的计算圆周率的方法，当时称之为“割圆术”。算法的英文名称为“algorithm”，来源于中世纪的拉丁语单词“algorism”，这个词

是为了纪念波斯数学家花拉子米(Al-Khwarizmi)，他最早在数学上提出了算法的概念，人们就把他名字的音译作为算法的名称，意思是“花拉子米提出的运算法则”。后来到了 18 世纪，这个词才演变为现在使用的“algorithm”。

关于算法的定义，也是形式多样，但基本意思都差不多。有的书把算法定义为“一系列的计算步骤，用来将输入数据转换成输出结果”；百度百科的定义是“解题方案的准确而完整的描述，是一系列解决问题的清晰指令，算法代表着用系统的方法描述解决问题的策略机制”；而维基百科对算法的定义则是“一个关于计算的有限长度的具体步骤，常用于计算、数据处理和自动推理，算法应包含清晰定义的指令用于计算函数”。

不管这些描述的侧重点和角度有什么不同，所有的算法都包括了几个共同的基本特征：输入、输出、明确性、有限性、有效性。

怎么来理解这几个特征呢？我们用“决定穿什么衣服”这个例子来做类比。一个算法是必须要有零个、一个或多个输入量的，这个输入就可以看作你衣橱里衣物的种类和数量、当天的天气情况、是上班还是休息等；输出量是算法计算的结果，每个算法应该有一个以上的输出量，也就是说你考虑了种种因素之后，必须决定选择哪套衣服作为当天的着装，哪怕是不穿衣服也是一种输出，尽管看起来很疯狂；算法的明确性是指算法的描述必须是无歧义的，比如这个决定着装的算法，对这个算法的描述就只能是关于如何着装，而不是早上吃什么；有限性是说算法的步骤必须是有限的，这个比较容易理解，无限的步骤显然永远也不可能执行完，从早上起床选择衣服，到天黑了还没选出来，这绝对是一件匪夷所思的事情，别人会认为你脑子出了问题；最后是有效性，这是说算法的操作和结果是能够实现的，如果最后得出的结果是你衣橱里没有的衣服，显然这是个无效的算法。

从古至今，涌现出的算法何止万万千千。在这其中，有精妙的算法，有一般的算法，也有拙劣的算法。比如，计算 1＋2＋3＋…，一直加到 100，最简单

粗暴的方法就是老老实实一个数一个数地算，当然用这样的“笨办法”也能算出来，还要保证计算过程中不能出错，但我们很难将这样的算法称为一个好算法。从 1 到 100 还可以用手工计算，但当长度增加到一千、一万、甚至是一亿，那要算到什么时候？所以，总有聪明的人，比如高斯，能够发现计算等差数列的好方法，不管有多少数，都能在很短时间内得到结果。具体到某一个算法是好是坏，这需要一个评判标准。一般来说，一方面是看算法运行的速度，也就是多长时间能够得到结果，当然是越快越好；另一方面是看算法运行需要占用的存储空间有多大，因为算法一般在计算机中运行，而计算机的内存空间是有限的。就像人的大脑，在同一时刻，大脑是在处理多个任务的，比如，某个码农可以一边敲代码一边喝咖啡，同时还在盘算下班后约几个死党找地方撸串。假如一个任务就把大脑撑满了，其他所有事，诸如动动手指、说话等都不能做了，那显然这样的算法是存在重大缺陷的。关于如何评价算法，这里先简单让大家有个概念，我们在后面还会专门叙述。

统治世界的十大算法

2014 年，加拿大未来主义者、生物伦理学家兼科学作家乔治·多沃斯基(George P. Dvorsky)推出了一个关于算法的排行榜，称为“统治世界的十大算法”，从计算机和互联网的角度评出了与人们生活息息相关的十个算法。高居榜首的就是 Google 公司创始人之一的拉里·佩奇(Larry Page)发明的 PageRank 算法。这是一个计算网页排名的算法，体现网页的相关性和重要性，Google 搜索给出的搜索结果就是基于这个算法，这可是 Google 公司发家致富的基础，其重要性可以与可口可乐的配方相提并论，在本书第 5 章还会对这个算法做详细介绍。

排名第二的是 Facebook 公司的 News Feed，这是一个新闻推送的算法，可

以看作美国版的“今日头条”。它根据人们的喜好不同将不同的内容推送给用户，很多美国人都承认“News Feed 是我们最喜欢浪费时间的地方”。而今，News Feed 的广告业务也为 Facebook 带来了每天近 4000 万美元的收入。

排名第三的是匹配算法，目前在线交友、婚恋匹配等已经成为动辄几十上百亿美元的产业，匹配算法在其中功不可没，在本书第 3 章将对匹配做详细介绍。

位于第四位的是与安全相关的数据采集、加解密算法。通过斯诺登，我们已经知道了美国国家安全局（NSA）及其小伙伴已经暗中监控了数百万对此一无所知的无辜民众。算法在这其中扮演了很不光彩的角色，当然，算法对此也很无奈，它也是做牛做马的悲情角色。

第五位是推荐算法，跟人们平日网络购物、看视频等活动息息相关。人们在淘宝或京东购买了衣物、书籍，或是在 Netflix、爱奇艺看了某个电影或连续剧，这些网站的推荐算法就会记录这些浏览和购买信息，然后利用这些信息“猜测”某个人的喜好，根据猜测结果为人们推荐商品。所以很多人会发现，当再次登录淘宝或京东时，系统会向你推荐和你曾经浏览过或买过的产品类似的产品，因为系统会认为你曾经买过，所以你再次购买的可能性会变大。读者可以自己判断这样的推荐算法是否会让你满意。在本书第 5 章也将对推荐算法做详细介绍。

排名第六的是 Google 公司的 Adwords，这是一种通过使用 Google 关键字广告来推广网站的算法。

排名第七和第八的分别是股票交易算法和 MP3 压缩算法。MP3 是一种音频压缩技术，于 1987 年由德国的一个研究组织 Fraunhofer-Gesellschaft 发明并标准化，它能将音频以 1∶10 甚至 1∶12 的压缩率压缩，从而大大降低了存储音频文件所需的容量。

排名第九的预测分析算法，这是一种让人细思极恐、会想起电影《少数

派报告》的技术。虽然这种技术还没有主宰我们的世界，但越来越多的警察机构正在使用这一预测技术。据说在2010年，美国孟菲斯市警察局通过使用IBM公司的预测分析软件CRUSH（Criminal Reduction Utilizing Statistical History），使当地恶性案件的发生率降低了超过30%，其中包括15%的暴力犯罪。很多其他国家和地区也开始关注这一技术，包括洛杉矶、圣克鲁斯、查尔斯顿等地方也开始了试点。这种算法结合了数据采集、统计分析等方法，发现所在城市的犯罪特点，并对可能出现的犯罪"热点"进行预测，从而可以"积极地配置资源和分配人手，提高人力物力的使用效率，提高公众安全"。

位居末位的是调音算法，这种算法最初是用于处理地震数据，后来人们意外发现在处理歌声或乐器的音色时有很好的效果。美国女歌手雪儿（Cher）演唱的"Believe"，被认为是第一首使用调音的流行歌曲。

对于"统治世界的十大算法"，也有很多人提出不同意见，认为这些算法都是在应用层面的，是建立在纯算法的基础上的，不是真正意义上发挥基础作用的算法，而且有为互联网公司打广告的嫌疑。因此，那些坚持"正统"的人就搞了一个"真正统治世界的十大算法"的榜单，包括排序算法、傅里叶变换算法、迪杰斯特拉算法、RSA算法、哈希算法等。这些算法的确在很多领域发挥着重要作用，其中一些我们会在后面的章节中介绍。

算法的载体——图灵和图灵机

计算机没什么用，它们只会告诉你答案。

——巴勃罗·毕加索

图灵的甜点

前文讲到了各种形形色色的算法，不管是应用类的还是基础类的，必然需要一个承载其运转的载体，就好比大脑是人类思维的载体。现实中算法的载体，就是各种计算机，算法与计算机之间是密不可分的关系，因此，讲算法必然要谈到图灵和图灵机，以及后来的图灵测试和人工智能。

艾伦·麦席森·图灵（Alan Mathison Turing）是英国数学家、逻辑学家，被称为计算机之父、人工智能之父。也许很多人是通过电影知道图灵这个人的。2015 年 2 月 23 日上午，第 87 届奥斯卡金像奖颁奖典礼在好莱坞杜比剧院举行，其中有一部名为《模仿游戏》（*The Imitation Game*）的传记电影斩获最佳改编剧本奖，以及包括最佳影片、最佳导演、最佳男主角、最佳女配角在内的 7 项提名。这部影片主要讲述了“计算机之父”图灵的传奇人生，着重展示了图灵帮助盟军破译德国加密机“英格玛”，从而扭转二战战局的经历。当然，作为艺术作品，这部电影也存在一些拔高或偏颇之处，但图灵的巨大贡献和功绩是为世人所公认的。

需要更正一下的是，最先为破解英格玛密码机做出重要贡献的是波兰人雷耶夫斯基（Marian Rejewski）。在这里，不得不感叹波兰人搞情报的能力。波兰人通过间谍活动得到了英格玛密码机的工作原理和内部构造的资料，以及德军对英格玛密码机的操作守则。根据这些资料和平日里截获的德军电文，雷耶夫斯基破译了德军日密钥的全部内容，可以说与德军中的报务员处在了完全对等的地位。在 1939 年德军入侵波兰前夕，波兰人将英格玛密码机的复制品和破解方法提供给了英国和法国。如果历史就这样不变，那其实就没有图灵什么事了。

在二战爆发前夕，为了战争的需要，德军又采取了很多措施来加强英格

玛密码机的安全性，使得波兰人前面的破解方法全部失效了。二战开始后，对英国来说，破解英格玛密码机的工作就显得非常重要了。此时，图灵大神开始登场。根据波兰人提供的资料，他意识到波兰人的破解方法过于依赖德国人操作方式上的漏洞。图灵的思路是，不管英格玛密码怎么复杂，它也是机器创造出来的，只有用机器才能战胜，人类的任务就变为找到并设计出这个机器的工作原理和优化机器进行的运算量。他想要做的是一种更纯粹和直接的暴力破解——发明一种更强大的机器去战胜英格玛密码机。

经过千难万险，图灵终于设计并制造出了破解英格玛密码机的机器——“图灵甜点”(Turing Bombe)。“图灵甜点”的样子如图 1.5 所示。这个机器里包含很多三个一组的转盘，每组转盘都相当于一台英格玛密码机，一台标准的“图灵甜点”共有 36 组这样的转盘。机器在进行暴力破解时，如果遇到可能的解，它就会停下来，以便工作人员进行记录；而它没有停下来的时候，人们就只能站在旁边等待。那么图灵为什么能够想到用机器来对付机器的思路呢？这其实还是跟“图灵机”的思想有关。



图 1.5　图灵甜点

图灵的精彩人生

图灵出生于英国一个没落的贵族家族，但他这一支属于旁系，不能继承爵位，没有领地，也不能继承多少财产。这样的旁系成员大多都成了神职人员或者英属殖民地的公务员，图灵的父亲就是一位服务于印度边远地区的基层行政人员。1912 年 6 月，图灵的父母回到伦敦休假期间，生下了图灵。在

他一岁的时候，因为父母要返回印度继续工作，图灵和他哥哥被交给居住在英格兰南岸的一对退役陆军上校夫妇照顾，这使得他的成长过程缺乏了必要的关爱，对他以后的性格和行为产生了比较大的影响。

后来，图灵的母亲回到英国，他才得以离开寄养家庭，跟母亲生活在一起。图灵在少年时代就展现出对数学的热爱和独特的创造能力。他在 13 岁那年考上了伦敦有名的舍伯恩（Sherborne）公学，接受了良好的中等教育。1927 年，在图灵 15 岁的时候，为了帮助母亲理解爱因斯坦的相对论，他为关于爱因斯坦一部讲述相对论的著作写了一个内容提要，表现出非凡的数学水平和理解力，但他也因为对文科很不重视，以至于文科成绩过于糟糕，经常遭到老师的责骂。也就是在舍伯恩公学的这段时间，图灵爱上了他的金发同学克里斯多夫・莫科姆（Christopher Morcom），此时，他才意识到自己是一个同性恋者。在那个时候，同性恋在英国算是一种严重的罪行，这为图灵以后悲惨的遭遇埋下了伏笔。很不幸，在中学毕业前的寒假，莫科姆感染了结核病并很快离世。这个变故差点使图灵崩溃，他自此以后变得更加独来独往，再也无法轻易与他人建立起亲密关系。

遭受这些打击后，跑步成了图灵释放压力和追求内心平静的唯一方式。即使成名以后，他也一直坚持这个习惯，在工作忙碌时，他也想方设法跑步，比如从家里跑到工作的地方，跑步到会议现场参加科学会议，这让他的同事们都觉得不可思议。图灵的传记作家安德烈・霍奇（Andrew Hodges）这样说道："最关键的是，他跑得比一些交通工具还要快。"更有意思的是，如果不是腿伤的原因，图灵还很有可能跑进 1948 年的伦敦奥运会。

图灵的个人马拉松最好成绩是 2 小时 46 分 03 秒，但因为持续训练和略显僵硬的跑步姿势，他遭遇了腿伤，以致错过了业余田协在 1948 年 6 月举办的伦敦奥运会马拉松选拔赛。那一年，伦敦奥运会马拉松比赛冠军的成绩是 2 小时 34 分 52 秒，英国选手拿到了银牌，而代表英国出战的第二好选手成绩也仅仅是 3 小时 09 分，远不如图灵。

在舍伯恩公学的最后一年，图灵获得了剑桥大学国王学院的奖学金，他于 1931 年进入该学院学习数学。图灵用获得的奖学金买了三本书，其中一本是约翰・冯・诺依曼（John von Neumann）所著的《量子力学的数学基础》，这本书中描述了在亚原子层面发生的事件如何受到统计概率的控制原理，这对图灵以后的学术道路产生了深远的影响。在此后很长一段时间，图灵都一直在钻研“人类大脑与确定性机器之间是否存在根本区别”这个问题，他后来逐步得出的结论是，人与机器之间的界线比他原来想象的更加模糊。

德国天才数学家大卫・希尔伯特（David Hilbert）在 1928 年的一场数学大会上，提出了关于任意数学形式系统的三个基本问题。前两个问题在之后三年不到的时间，就被来自奥地利的数学家、时年 25 岁的库尔特・哥德尔（Kurt Gödel）解决了。希尔伯特三个问题只剩下最后一个：“系统是可判定的吗？有没有可以判定特定命题是否可证明的方法？会不会出现某些陈述存在不可判定状态的可能性?”希尔伯特本人把这第三个问题称为“判定问题”。1935 年，在剑桥大学数学教授麦克斯・纽曼（Max Newman）开设的数理逻辑课上，图灵第一次听到了希尔伯特的判定问题。纽曼教授根据自己的理解给出了对判定问题的表述：“有没有一种‘机械方式’可以用于判定某个逻辑命题是否可证?”图灵开始对判定问题产生了兴趣，他非常喜欢纽曼教授提出的“机械方式”的概念。图灵敏锐地意识到，解决这个问题的关键，在于对这种“机械方式”的严格定义。可以说，这个概念在一定程度上促进了以后“图灵机”思想的诞生。

跑步后的灵感——图灵机

1935 年夏季的一天傍晚，图灵又像往常一样沿着伊利河畔跑步。在以每英里[1] 4 分半的速度跑完了大约 6.2 英里的路程后，图灵躺在格兰切斯特

① 1 英里＝1.609 344km。

草甸的苹果树下，一边享受落日的余晖，一边潜心思考着一个困惑已久的问题。关于“机械方式”的定义，他还是觉得“用机器来完成计算”是一个可行的方式。其实，“用机器来代替人工进行计算”的想法并不新鲜。17世纪，德国哲学家、数学家莱布尼茨（Gottfried Wilhelm Leibniz）就曾经设想过用机械计算来代替哲学家的思考。图灵的设想与先贤们略有不同。在他的设想里：这个机器必须足够简单，要简单到能够用逻辑清晰的公式来描述它的行为，也要能够造出实物；它又要做到足够复杂，要具备能够完成所有机械能完成的运算。图灵打算构造的是一种能够产生复杂行为的简单机器。

不知道是不是跑步带来的效果，图灵那一天的思路特别清晰，以前一些卡顿的地方也都豁然开朗。很快，在他的头脑中勾勒出了一种机器的模型，这正是他梦寐以求的样子。图灵一跃而起，兴奋地挥舞着双手，虽然他并没有被苹果掉下来砸在头上，但图灵感觉到确实是有如神助。

图灵回到学校，洗了一个舒舒服服的热水澡，平复了一下心情，然后马上到工作室记下了当时的思路。图灵的基本思想是用机器来模拟人类用纸笔进行数学计算的过程，他假想出一台结构非常简单的机器，这种机器由两部分组成：一个读写头，一条无限长的纸带，纸带分成一个一个小方格，每个方格中只能有0和1两种符号。读写头可以在纸带上移动，读取或改变纸带格子上的信息，但每次只能对着纸带上的一个格子。读写头可以处于不同的状态，但状态的数目是由状态转移表确定的，并且是有限的。每一时刻，读写头从纸带上读入一个方格的信息，然后根据状态转移表查找内部的固定程序，根据程序输出信息到纸带的方格上，并转换内部状态，然后移动。在读写头的所有状态中，有一个特殊的“停机”状态，读写头如果处于停机状态，就会停止工作，否则它会一直运行下去。实际上，这台机器最关键的地方就在于状态转移表，它指示读写头的状态及其相应格子里的信息如何变化，可以说，状态转移表描述了这台机器执行的算法。需要注意的是，图灵机并不是一台真

实的机器，而是一种抽象的计算模型。虽然是抽象模型，仍然有人能够把图灵机制作出来，图 1.6 就是乐高公司制造的一台玩具图灵机。

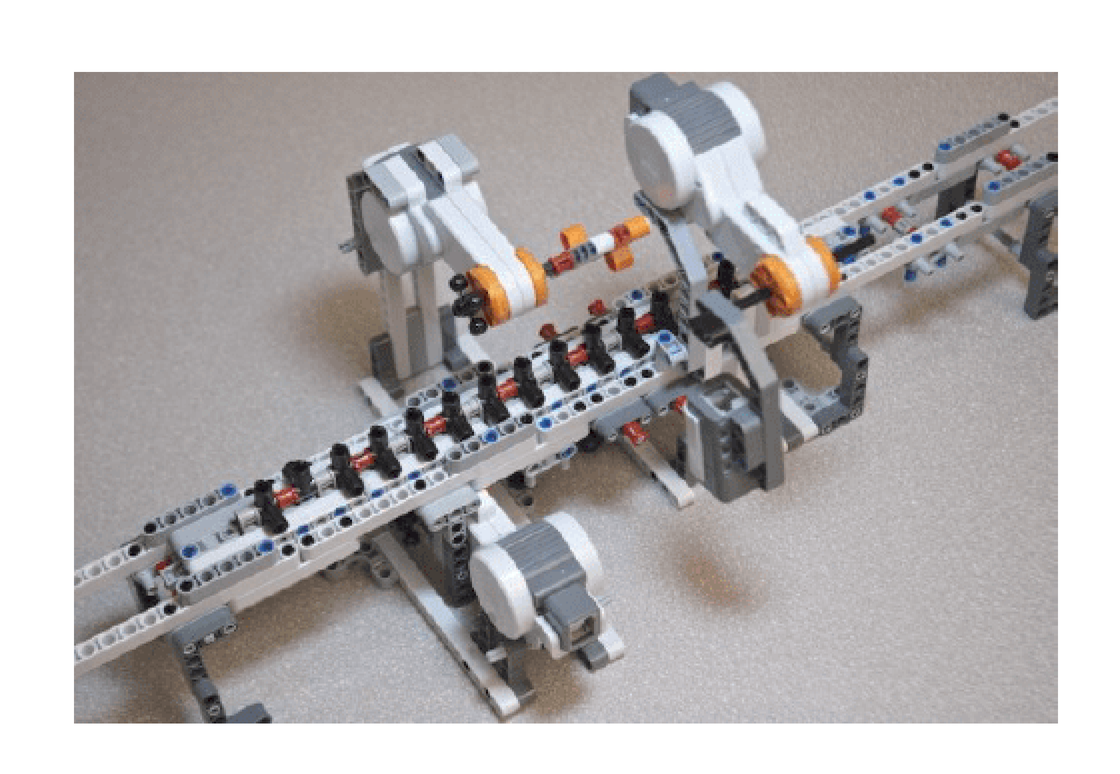


图 1.6　乐高玩具版图灵机

实际上，这台机器的运转过程跟人类笔算乘法的过程非常类似：在每个时刻，我们的注意力都集中在笔尖上，根据眼睛看到的信息移动笔尖，在纸上写下符号，而指示我们怎么写、写什么的则是我们早已熟悉的运算法则。在这里，计算用的纸张就是纸带，读写头就是人和他手中的笔，状态转移表就是乘/加法规则表，读写头的状态就是大脑当前的精神状态。

1936 年 9 月，图灵应邀到美国普林斯顿大学高级研究院学习，在这里，他结识了数学家阿隆佐·丘奇（Alonzo Church）并同他一起工作。丘奇在判定问题上的研究一直领先于图灵，但他胸怀坦荡，乐于提携后辈。图灵在 1937 年将自己的研究成果写成论文《论可计算数及其在判定问题上的应用》，并将论文副本送交 6 位知名学者审阅。只有丘奇认真阅读了图灵的论文，他非常欣赏图灵的工作，不仅给出很多好的建议，还将图灵设计的计算机器命名为“图灵机”。从此，“图灵机”横空出世，也标志着这位年仅 24 岁的年轻人的名字将被永远刻在数字时代最为重要的里程碑上。

图灵机的思想对于后来现代计算机的发展影响非常深远。共同创立 TCP/IP 协议的“互联网之父”温特·瑟夫（Vint Cerf）与罗伯特·卡恩

(Robert E. Kahn)在一篇纪念图灵的文章里写道：

> 如果他能够活到今年，不知他看到今日的景象会有如何的感想，又能为我们提出哪些值得思考的问题。我花了一辈子研究计算机和网络，可是时时还会期盼图灵能够在身边回答我的疑问。相信如果有他在，很多问题都能够迎刃而解。

从图灵测试到人工智能

> 计算机会不会思考这个问题就像问潜水艇会不会游泳一样。
>
> ——迪杰斯特拉，图灵奖获得者

你在跟谁对话

1950 年，我们的图灵大神又发表了一篇划时代的论文《计算机械与智力》，在这篇文章里讨论了创造真正的智能机器的可能性。他在文章的开篇就说道："我建议大家考虑这样一个问题：'机器能思考吗？'"。这个问题非常不好回答，并不是简单地说"能"或者"不能"就可以了。这个问题真正的难点在于怎么精确定义"思考"这个概念，如果纠缠于这个定义问题，很可能陷入哲学讨论的"泥沼"，偏离了问题的本来目的。很显然，图灵也看到了这一点，于是他很"机智"地绕过了这个定义，提出了一个可操作性很强的标准：如果一台机器表现得和一个会思考的人类一样，旁人无法区分，那我们就可以当作它是在"思考"。这就是著名的"图灵测试"，或者说，是一台机器模仿人思考的过程，模仿得越逼真，就说明它越"会"思考。而机器所有的这些模仿行为和思考过程，实质是算法，是算法赋予了机器智能，所以说图灵测试的

真正对象其实是算法。

到了1952年，图灵提出了一个具体的测试方法：让机器与人类进行对话，当然对话的人看不到他的对手（关在小黑屋里），只能根据对话过程来判断，如果有足够多的人（图灵设定的标准是30%）认为跟自己说话的是人类而不是机器，那么机器就成功通过图灵测试。从这个方法可以看出来，图灵测试的真正核心是在于“机器是否能够在智力行为上表现得和人类无法区分”，至于具体的形式是对话还是唱歌，或是别的什么都无所谓。图灵测试是人工智能哲学方面第一个严肃提案，而这个方法也在后世被发扬光大，成为图灵测试的唯一方法。

图灵本人确信机器是可以思考的，他还预言“在20世纪末，一定会有计算机通过图灵测试”“到2000年将会出现足够好的计算机，能够在长达5分钟的提问中全部回答正确”。著名的科普网站——科学松鼠会，曾经发布过一张名为“第一次成功的克林贡①图灵测试”的漫画，生动地描绘了图灵测试的过程（图1.7）。

公认的计算机首次通过图灵测试的时间是2014年6月7日——图灵逝世60周年的纪念日，比图灵预料的稍微晚了一点点。在英国皇家学会举行的“2014图灵测试大会”上，一个由俄罗斯团队开发的智能聊天程序，模仿成一个13岁的名叫“尤金·古斯特曼”（Eugene Goostman）的乌克兰男孩，通过了图灵测试。这次大会共有5个聊天机器人参赛，其中，尤金被33%的评委判定为人类，超过了图灵设置的30%的标准。

2014年5月，微软公司Bing搜索中国团队发布了智能聊天机器人“微软小冰”，开始了历史上最大规模的图灵测试。据微软高级研究员和科学家王

① 克林贡，英语Klingons，是著名的科幻影视《星际迷航》中一个好战的外星种族，几乎人人都是战士。

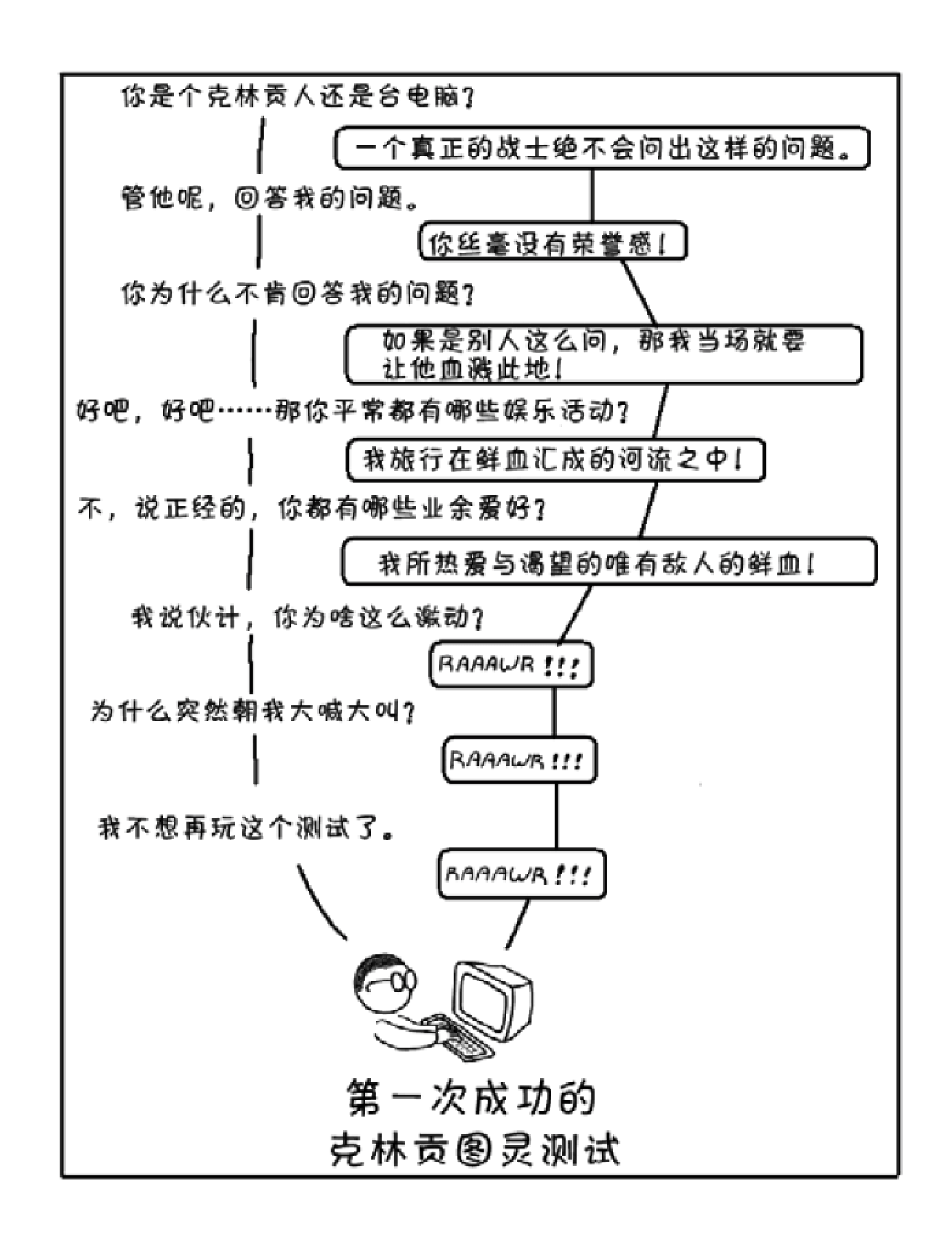


图 1.7　图灵测试漫画

永东披露，微软小冰可以像人一样回答问题，可以就任何话题与人们进行交流，还能提出问题甚至思考问题，比如，如果用户发出一张手指被割伤的图片，微软小冰会问是不是做饭弄伤的，还会问伤口疼不疼。微软小冰自上线以来，目前已经进化到第四代，每天约有数百万人通过各种平台与微软小冰对话，当然人们最喜欢的还是从各方面调戏微软小冰，互联网上有非常多调戏微软小冰的对话截图（图 1.8）。微软小冰也很有意思，她如果碰到自己不了解的东西，会尝试进行掩饰，如果实在掩饰不了，她还会像人一样恼羞成怒。据美国科技媒体 GeekWire 报道，已经有数百万中国用户向微软小冰表白过，大约 25％或 1000 万人在使用这项服务时说过“我爱你”。



图 1.8　微软小冰对话截图

2016 年 3 月,清华大学语音与语言实验中心(CSLT)网站宣布,他们开发的做诗机器人“薇薇”通过了来自社科院等不同单位唐诗专家的评定。在“薇薇”所创作的诗词中,有 31%被专家认定为人类创作的。下面来欣赏几首“薇薇”的作品,是不是难以想象它们竟然出自一台冷冰冰的机器?

早　梅

春信香深雪,
冰肌瘦骨绝。
梅花不可知,
何处东风约。

海　棠　花

红霞淡艳媚妆水,
万朵千峰映碧垂。

一夜东风吹雨过，

满城春色在天辉。

镜

照影金精映，

钗头角黍青。

白发红袖下，

明月满庭清。

值得一提的是，2017 年 5 月，微软公司和湛庐文化合作推出了微软小冰原创诗集《阳光失了玻璃窗》，这也是人类历史上第一部 100% 由人工智能创作的诗集。

从小男孩尤金，到微软小冰，再到薇薇，看起来，现代软硬件的进步确实如图灵所预测那样，有了天翻地覆的改变，通过图灵测试仿佛也是一件比较轻松的事情。但是，通过了图灵测试，就说明具有人工智能了么？这个问题同样难以有确定的答案，也许这仅仅只是一个开始。

人工智能：天使还是恶魔

到目前为止，人们理想中的人工智能，仍然停留在影视作品里。久远一点的有阿诺德·施瓦辛格（Arnold Schwarzenegger）主演的《终结者》（*The Terminator*，1984）、史蒂文·斯皮尔伯格（Steven Allan Spielberg）执导的《人工智能》（AI，2001），近一点的则有在第 88 届奥斯卡金像奖上斩获“最佳视觉效果奖”的《机械姬》（*Ex Machina*，2015），以及美剧《西部世界》（*Westworld*，2016）、《真实的人类》（*Humans*，2015）等。《机械姬》中的智能机器人艾娃（图 1.9），其创造者是某知名搜索引擎公司的老板、智商颇高的 IT 大牛，而貌似中大奖与亿万富翁老板共度周末的程序员加利，既没有享受到美女如云

的豪华派对,也没有到什么度假胜地,真正的使命却是悲催地“加班”,对智能机器人艾娃进行图灵测试。虽然,码农们更喜欢这种方式的超级大奖,这比酒精、美女更能让他们兴奋。加利也确实是真码农,完全没有对被“骗”有任何不满,而是兴致勃勃地踏入老板设下的局(观众总是喜欢看到码农的人生被如此安排)。值得注意的是,这部电影里进行的并不是传统意义上的图灵测试,而是该测试的升级版。

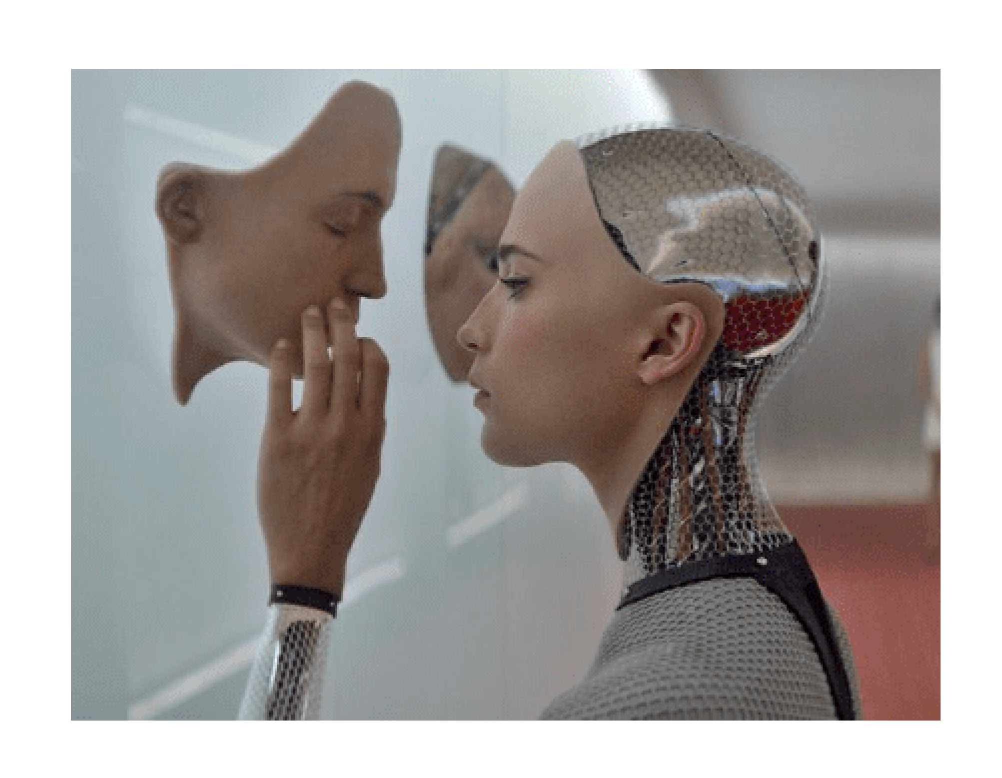


图 1.9 《机械姬》剧照

在图灵大神最初设定的测试环境里,被测试的对象是被关在一个黑屋子里,测试人看不到测试对象,只能根据对话来判断对方的身份。老实讲,在这种情况下,更多的只能判断测试对象的表现是不是更接近人类,说白了,就是像不像人说的话或反应。但要判断测试对象是否具有智能,就显得捉襟见肘了。当然,以图灵大神当时身处的时代,做到这一点已经难能可贵了。在《机械姬》里,加利与艾娃之间虽然隔着一层玻璃,但都可以看到对方,并且艾娃的机器身体并未增加多余的掩饰。其实就对话而言,艾娃要想通过图灵测试简直就是轻而易举的事,但大 Boss 的真正意图是要比图灵测试更进一步——确认艾娃是否具有真正的“智能”。在电影的设定里,艾娃是具有很强“智能”的,甚至达到了人们无法想象的程度,否则也不会把我们的码农男主角迷得晕头转向。当然,影片的最后也难免落入俗套,人工智能进化出了反

叛意识，想要挣脱人类的桎梏，甚至完成了反杀，而倒霉的男主则不幸成为被殃及的池鱼。

虽然在影视作品中高度发达的人工智能目前还没有实现，但自从图灵测试以后，随着计算机技术和人工智能算法的不断发展，特别在近几年，人工智能、机器人、智能家居等早已成为曝光度最高的词汇。不少专家大胆预测，互联网的下一波浪潮必然出现在人工智能领域。不过，有关人工智能的争论也愈演愈烈，力挺派和质疑派之间的争论从没停止。

力挺派认为人工智能是好事，绝对是未来的发展方向，会将人类从繁重单调的劳动中解放出来，会给人类社会带来美好的未来，会使人类发展往前迈出一大步等等。《连线》(*Wired*)杂志创始主编凯文·凯利(Kevin Kelly)在其力作《必然》中就表现出了坚定的支持："未来的人工智能网络(主要包含算法)将会成为'如同电力一样无处不在、暗藏不现的低水平持续存在'。"近年来算法的快速发展，特别是神经网络(Neural Networks)算法在近年的高速发展，给了力挺派很大信心，他们坚定地认为 21 世纪人类最重大的发明一定是"真正的人工智能"，瑞士人工智能实验室的科学事务主管施米德胡贝(Jürgen Schmidhuber)是其中的代表人物之一。他在德国慕尼黑工业大学学习期间先后获得计算机科学的学士和博士学位，时间分别是 1987 年和 1991 年。从 1987 年开始，他就是"自我改进式通用问题求解算法"领域的领头羊，从 1991 年获得博士学位后，又成为了深度学习神经网络算法领域的开拓者。他的研究团队提出了一种名为递归神经网络 RNN 的技术以及一些相关的变种，这些神经网络跟最初的相比，已经扩展了太多，可以做到成百上千层，每一层都是无数神经元处理单元的集合。这种新的神经网络层数更多、功能更强，理论上能够运行任意算法，或实现程序与环境之间的互动。施米德胡贝博士开发的这些新技术革新了手势识别、机器翻译、语音识别、图片注释等技术，已经被包括谷歌、微软、百度、IBM 在内的很多公司应用。有意

思的是，RNN 最初提出来的时候并不被看好。1995 年，他们提出一种能够学习长期依赖信息的 RNN，命名为 LSTM 网络（Long Short Term），还为此专门写了一篇论文，没想到居然被著名的机器学习学术会议 NIPS（*Conference and Workshop on Neural Information Processing Systems*）拒之门外。

施米德胡贝博士认为，以现在的神经网络为基础，不需要很多年，就可以实现跟某些动物一样聪明的人工智能，“一旦我们实现了动物水平的人工智能，几年或者几十年以后我们就可以拥有人类水平的人工智能了，那是真正没有局限的应用，所有的商业都会改变，整个文明都会改变，一切都会为之改变。”同时，他还认为，人工智能并不会对人类生死存亡带来新威胁，反而是核武器的威胁更大一些。

这里简单介绍一下神经网络算法。神经网络算法是从仿生学的角度出发，通过模仿人类大脑神经突触连接的结构以及行为特征，进行分布式并行信息处理的一种算法模型，它的目的是希望能够建立类似人脑的处理信息的机制。通俗地讲，人类的大脑有超过 100 亿个神经元，平均每个神经元都和另外 10 000 个神经元连接在一起。这些神经元中有些是与接收信息相关，比如听到声音、看到各种事物、感受到疼痛、触碰到不同物品的感觉等；另一些神经元与控制相关，比如控制肌肉做出各种动作等；除了这两种，剩下的大多数神经元隐藏在这两者之间，目前对它们还不够了解，但这部分神经元处理的是最复杂的思考过程。所有的神经元都通过改变连接强度进行学习，连接强度决定了神经元之间彼此影响力的大小。神经网络就是用计算单元来模拟神经元，并把这些计算单元连接起来形成网络。

当然，质疑派一直对此心存疑虑，他们提出了三个问题：第一，人工智能是否具有真正的“理解力”？第二，人工智能说到底是“人工”，错误不可避免。第三，人工智能自我进化不可控制。

关于理解力的问题，这涉及一个著名的思想实验——玛丽黑白色彩实验。研究色彩的女科学家玛丽从一出生就被关在一间房子里，这个房间里所有的布置和各种设施，包括玛丽穿的衣物，都只有黑与白两种颜色，这里就是她所有的生活空间。在这里，玛丽获得了所有关于色彩的物理知识（光学知识或神经学知识等），但没有亲眼看到过这些颜色。直到有一天，玛丽终于获准走出房屋，看到了真实世界中的各种色彩，比如，她看到了红色的西红柿。那么，问题来了：玛丽是否获得了更多的知识？对于玛丽看到真实西红柿的反应，一般人的直觉应该是，玛丽看到西红柿应该会获得某种直观的感受，也就是说，玛丽可能会惊讶：原来真正的西红柿（红色）是长这样啊！如果大家都承认这种情况会出现，那是不是就说明了物理知识并不是所有事实的知识，它并不完备，因为如果物理知识是完备的，则玛丽不会出现惊讶的感觉。简单地讲，这个结论是说明了知识和真实感受之间的区别，比如别人告诉你挨揍会痛，与真正被揍感觉到的痛是两个概念。

同样，具有人工智能的机器或算法，所拥有的知识都是这类物理知识，即使它能表现出类似人类的理解力，但是，它真正理解吗？从机器嘴里蹦出来的“我认为、我感觉”这些话，是它真正的想法吗？

人工智能说到底还是人弄出来的，人类本身就很难避免错误，再面对人工智能这么复杂的系统，能保证不出问题吗？2011 年合众国际社报道，美国加州大学的生物学家迈克尔·艾森在做研究时，需要从一本名叫《苍蝇的成长过程》（*The Making of a Fly*）的书中查找资料。很自然地，他在亚马逊网站上进行搜索，不料搜索结果让他大吃一惊，这本书的单价居然是 170 万美元，“肯定是亚马逊弄错了，我等等吧”，艾森当时这么想。就这样，艾森每天都登录亚马逊，查看这个错误定价是否已被修改。然而，令艾森意外的是，这个问题不但没有被修正，反而价格越来越高，在一周之内，居然飙升到了 2369 万美元一本。这是一本什么神书！？后来，亚马逊终于发现了这个问

题，把价格改回了正常水平。但这是怎么回事儿呢？原来问题出在算法上。亚马逊的图书定价是算法决定的，算法会根据其他卖家的定价、供求关系等因素来实时调整图书的价格。但当初设计定价算法的码农没有注意到其中一个 Bug——缺少一个价格上限。结果，机缘巧合之下，两个网站的定价算法在这本书上开始了死磕：你涨一毛，我就要比你多 5 分。最后的结果导致了天价图书的出现。质疑派难免会感到悲观：你看，就连给书定价这么一个简单的智能算法都会出现问题，我们怎么能够对人工智能算法有信心？

同时，质疑派最大的担忧还来自于人工智能的进化是否能受到有效的监督和约束。《连线》杂志在 2014 年 6 月刊出的一篇文章 *Why We Need to Tame Our Algorithms Like Dogs*（我们为什么要像驯狗那样驯化算法）中谈到："人类目前正在与另外一种我们之外的物种共生在一起，与犬科动物相比，它更加危险也更有威力：这就是算法。"

这样的说法似乎也不无道理。2016 年 11 月 17 日，在深圳举办的第十八届中国国际高新技术成果交易会上，一台名为"小胖"的机器人在没有收到任何指令的前提下，突然暴打所在展台的玻璃墙，展台破坏严重，玻璃碎片四处飞溅，一名运气不太好的参观者被玻璃碎片划伤，最后被送上 120 急救车，如图 1.10 所示。虽然事后涉事参展商负责人发布消息说，机器人异常行为是由于工作人员的误操作导致的。但是真相如何，众说纷纭。2017 年 6 月，《大西洋月刊》网站也报道了一则机器人不受控的例子：Facebook 在实际中让两个 AI 聊天机器人对话，发现机器人竟逐渐发展出人类无法理解的独特语言，研究人员不得不对其进行人工干预。AI 自行升级的速度超出预期，让人不禁担心《银翼杀手》中的恐怖场景会在现实生活中出现。

为了确保人类自身的安全，20 世纪最伟大科幻作家之一的伊萨克 · 阿西莫夫于 1942 年在其短篇作品 *Runaround* 中，首次提出了机器人三大法则，即

（1）机器人不得伤害人类，或因不作为使人类受到伤害；

（2）除非违背第（1）法则，机器人必须服从人类的命令；

（3）在不违背第（2）法则的情况下，机器人必须保护自己。

后来，阿西莫夫还加入了一条第（0）条法则，即

（0）机器人不得伤害人类整体，或因不作为使人类整体受到伤害。





图1.10　涉事机器人“小胖”和事故现场

阿西莫夫规定，这几条法则都必须植入机器人智能算法的底层，所有机器人都必须遵守。但是，别的不说，单单这几条法则，本身就存在漏洞、缺陷和模糊之处，比如“人”的定义、“机器人”的定义，都还在争论之中，没有形成定论。这些模糊或不完整的信息，完全有可能导致人工智能的行为不可控，打破这些所谓的“铁律”。艾娃的算法里是写入了不能伤害人类的指令，但艾娃在自主进化过程中不但进化出控制电力的能力，也完全摆脱了不能伤害人类的桎梏，所谓的铁律在进化的力量面前不堪一击。美国斯坦福大学的生物学家黛博拉·戈登（Deborah Gordon）在接受《量子杂志》的采访时曾谈到：“进化可能会在不同系统中创造出用来解决相同问题的不同算法。”这充分说明了进化过程和结果的不可控。例如，制造出100个一模一样的艾娃，经过相同的进化历程，可能有些艾娃会遵守指令老老实实待在小黑屋；有些可能

会请求大 Boss 带她出去看看外面的风景;而有些可能会像影片中那样,进化出暴力的解决方案。

再回想一下前面提到的让人工智能自动纠错、进化的神经网络算法,超大的规模已经远远超过人所能控制的范围,有谁能保证这其中不是危机四伏?有可能一次进化的微小错误,就会导致人类巨大的灾难,就像那只南美洲亚马逊河流域热带雨林中的蝴蝶,只不过轻轻扇动一下翅膀,两周后就会引起美国得克萨斯州的一场龙卷风。

加拿大学者马歇尔·麦克卢汉(Marshall McLuhan)在其著作《理解媒介》中写道:“起初,我们塑造了工具,最后工具又反过来塑造我们。”现在,我们正走在塑造人工智能的路上,未来,人工智能会怎样反塑人类,是用一种温和渐进的方式,还是激烈突变的方式?谁也不知道。

算法的复杂性

复杂性杀死一切。它把程序员的生活给搞砸了,它令产品难以规划、创建和测试,带来了安全挑战,并导致最终用户和管理员沮丧不已。

——雷·奥兹 Microsoft

耗时又耗力的算法

有些算法很简单,但更多的是复杂的算法。高斯小时候计算等差数列的算法,大概需要四分之一张 A4 纸,一支铅笔,在 5 分钟之内就可以得到结果。较复杂的算法,可以回想一下我们在前面介绍过的,图灵为了破解德军

的英格玛密码机，制造了一台名叫"图灵甜点"的机器来对抗。"图灵甜点"在进行暴力破解时，如果遇到可能的解，它就会停下来，以便工作人员进行记录；而它没有停下来的时候，人们就只能站在旁边等待。这个时间可长可短，短的话可能几十分钟，长的话可能得好几个小时。这是图灵破解算法的一个大概的执行时间长度。

更复杂一些的问题呢？

有一个名叫"$3x+1$"的猜想，它的内容很有趣："任取一个自然数，如果：(a) 它是偶数的话，就把它除以 2；(b) 如果它是奇数的话，就把它乘上 3 再加 1；这样我们就得到一个新的自然数，对这个新的自然数再继续重复这两步变换，会得到一串自然数，最后一个自然数一定是 1。"

我们用自然数 4 来试一下。4 是偶数，把它除以 2，我们得到 2；2 仍然是偶数，再把它除以 2，确实得到的是 1。我们再找个大点的数，比如 7。7 是奇数，把它乘以 3 再加 1，得到 22；22 是偶数，把它除以 2 得到 11；11 是奇数，……，最后我们得到的一串数是 7→22→11→34→17→52→26→13→40→20→10→5→16→8→4→2→1。经过 16 步变换，我们最后真的得到了 1。如果你有足够的时间和兴趣，还可以尝试一些更大的数，试一试看最后是不是都回到 1。

其实，这个猜想在西方经常被称为西拉古斯(Syracuse)猜想，因为据说这个问题是 20 世纪 50 年代在美国西拉古斯大学(Syracuse University，也译为雪城大学)提出来并开始研究的。耶鲁大学教授、日本数学家角谷静夫将这个问题带到日本，因此在东方，这个问题也称为角谷猜想。此外，它还有一大堆各式各样的名字，大都与研究它的数学家有关，比如克拉兹(Lothar Collatz，德国数学家)问题、乌拉姆问题(Stanislaw Ulam，波兰裔美国数学家)等等。后来，大家为了方便，就把它简称为"$3x+1$"猜想。

据角谷教授介绍，当时有整整一个月的时间，整个耶鲁大学的人都在试

图解决这个问题，但是不幸的是，没有取得任何结果。后来在芝加哥大学也上演了同样的场景，同样是无疾而终。这个问题让人如此疯狂，以至于有人甚至猜测，这个问题是苏联克格勃的阴谋，真正的目的是要阻碍美国数学的发展。这样的猜测令角谷教授也忍俊不禁，他对克格勃是否有如此远大的数学眼光表示怀疑，不过他也表示，对于这种形式如此简单，解决起来却如此困难的问题，实在是可遇不可求，拥有非凡的魅力。

现在我们用图灵机来尝试一下是否能够解决这个问题。假设我们现在有一台制造出来的图灵机，有无限长的纸带，我们将“$3x+1$”猜想写成一个简单的算法程序，存到图灵机里，让它帮我们做一件事情：找出一个不满足猜想的数并输出到纸带上。于是，算法开始指挥图灵机的读写头移到纸带的第一格，读入第一个数 1，检验是不是符合要求，然后读入第二个数，根据规则进行判断，然后做相应的操作，并将中间过程所得到的数记录在纸带上。读写头就这样不断移来移去，很长时间都不会停下来，如果安排一个 20 岁的年轻学生每天来查看一下是否找到了不满足规则的数，也就是读写头是否会停下来，以读写头的速度，估计等他到了耄耋之年，都不会停下来。

然而，得益于现代计算机技术的发展，计算速度有了飞速提高，使得我们在较短的时间内就可以算到很大的数。目前，据说有人已验证过的最大的数是 $2^{50}\times100=$ 112 589 990 684 262 400，从 1 到这个巨大的数，通过变换最后都会回归到 1，无一例外。图 1.11 展示了验算从 1～9999 的每个数需要花费的时间。

验算这个猜想的时候，除了那个不知疲倦、勤勤恳恳来回奔波的读写头，另外一个需要注意的就是那条长长的纸带。在这条纸带上，记录着所有被验算的数，以及在验算过程产生的中间结果，也就是说，像上面验算“7”的例子一样，每个数后面都跟着一串数字序列。我们无法想象这条纸带究竟会有多长，也许从地球到月球（约 38.4 万公里）这么长的距离也是不够的吧。

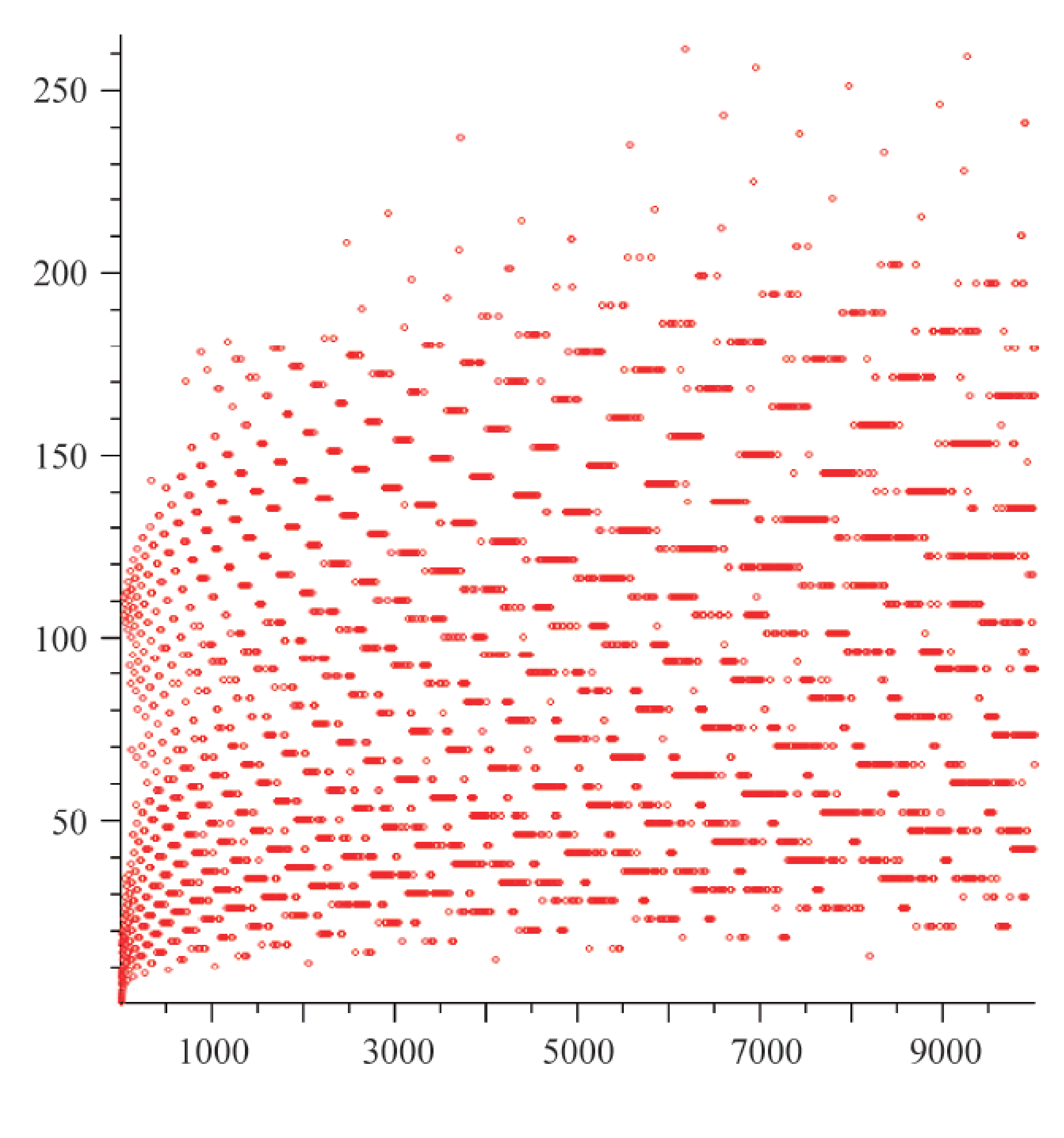


图 1.11　验证“$3x+1$”猜想的时间开销图

度量算法的两大基准

通过上面的例子，大家也许已经发现，与算法性能有密切相关的是两个方面，一是多久可以得到最终的结果，也就是上面所讲的读写头多久能停下来，给出一个不能变换到 1 的自然数；二是需要多大容量来支持，也就是上面所讲的纸带有多长，纸带越长，能容纳的结果越多。的确，从时间和空间这两个维度上可以衡量算法的优劣程度。事实上，现在人们也正是这么做的，评估方法的正式名称，称为时间复杂度和空间复杂度。

从字面上看，时间复杂度和空间复杂度并不难理解。时间复杂度用来衡量算法执行所花费的时间，是表达时间变化规律的概念。算法是由计算机语句（程序）来描述的，因此一个算法执行所花费的时间与需要执行的程序的语

句多少成正比，需要执行的语句越多，算法就越耗时。人们给算法中语句的执行次数起了一个名字，称为时间频度，用符号 $T(n)$ 来表示。时间频度中的 n 称为问题的规模，显然，n 越大，时间频度越大。一般来说，算法中语句的执行次数是 n 的函数，也就是时间频度，如果存在某个函数 $f(n)$，使得在 n 趋近于无穷大的时候，$T(n)$ 与 $f(n)$ 比值的极限是不为零的常数，就称 $f(n)$ 是 $T(n)$ 的同量级函数。时间复杂度用大写的英文字母 O 表示，也就是 $T(n)=O(f(n))$。一般情况下，标记时间复杂度时，只取 $f(n)$ 的最高次项，也就是运算次数受影响最大的那一项，不包括该项的系数和低阶项。比如，对于 $T(n)=n^2+7n$ 和 $T(n)=5n^2+3n+5$，它们的时间频度不同，但时间复杂度都是 $O(n^2)$。

相似地，空间复杂度定义为运行完算法（程序）需要的存储空间的大小，包括不变空间和可变空间两部分，它也是规模 n 的函数，同样用 O 来表示。

下面，我们以一个计算最短路径的算法为例，看看时间复杂度和空间复杂度是怎么得到的。计算机科学里有一个著名的 Dijkstra 算法，中文发音称为迪杰斯特拉算法。这个算法是用它的发明人艾兹赫尔·韦伯·迪杰斯特拉（Edsger Wybe Dijkstra）的名字命名。迪杰斯特拉来自荷兰，他的贡献非常多，覆盖了很多领域，是计算机科学的奠基人之一。

迪杰斯特拉算法是经典的单源最短路径算法，也就是说，计算从一个节点出发，到其他所有节点的最短路径。迪杰斯特拉算法在生活中的应用场景很多，比如网络路由查找、城市规划、地质勘探、车辆导航、物流配送等。以下是这个算法的基本思想。

考虑一个路径搜索的例子，以图 1.12 为基础，计算从清华大学出发，分别到达鸟巢、国家图书馆、鼓楼、北京西站和东直门的最短路径。如果某两个地点之间有直接相连的边，说明这两个地点直接可达，这条边上的数字表示经过这条边所需要花费的代价，可以是时间、金钱等。如果某两个地点之间

没有直接相连的边，就表示这两个地点不能直达，必须要通过别的地点，比如清华大学到鼓楼不能直达，代价为无穷大，只能通过鸟巢或国家图书馆过去。下面开始计算：

（1）选定清华大学为起点，则此时地点集合 V＝{清华大学}，路径集合 E＝{}。

（2）从清华大学出发，开始搜索，发现可以到达鸟巢和国家图书馆，代价分别是 6 和 3，因为 3 小于 6，所以选定代价为 3 的这条边，将新的地点和边加入地点集合和边集合，此时 V＝{清华大学，国家图书馆}，路径集合 E＝{清华大学→国家图书馆＝3}。

（3）从国家图书馆这个顶点开始搜索，可以发现从清华大学经过国家图书馆到达鸟巢、鼓楼和北京西站的代价分别是 5、6、7（注意从清华大学直接到鸟巢的代价是 6，大于 5），因此选取代价为 5 的边，此时 V＝{清华大学，国家图书馆，鸟巢}，路径集合 E＝{清华大学→国家图书馆＝3，清华大学→国家图书馆→鸟巢＝5}。

（4）类似地，从鸟巢开始搜索，得到的 V＝{清华大学，国家图书馆，鸟巢，鼓楼}，路径集合 E＝{清华大学→国家图书馆＝3，清华大学→国家图书馆→鸟巢＝5，清华大学→国家图书馆→鼓楼＝6}。

（5）以此类推，最后得到 V＝{清华大学，国家图书馆，鸟巢，鼓楼，北京西站，东直门}，路径集合 E＝{清华大学→国家图书馆＝3，清华大学→国家图书馆→鸟巢＝5，清华大学→国家图书馆→鼓楼＝6，清华大学→国家图书馆→北京西站＝7，清华大学→国家图书馆→鼓楼→东直门＝9}。

此时，所有的地点都加入了地点集合，算法结束，从清华大学出发分别到达 5 个目的地的最短路径就这样出现了，如图 1.12 右图中实线路径所示。从上面的过程可以看出，迪杰斯特拉算法是一种按路径长度递增的次序来找出最短路径的算法。

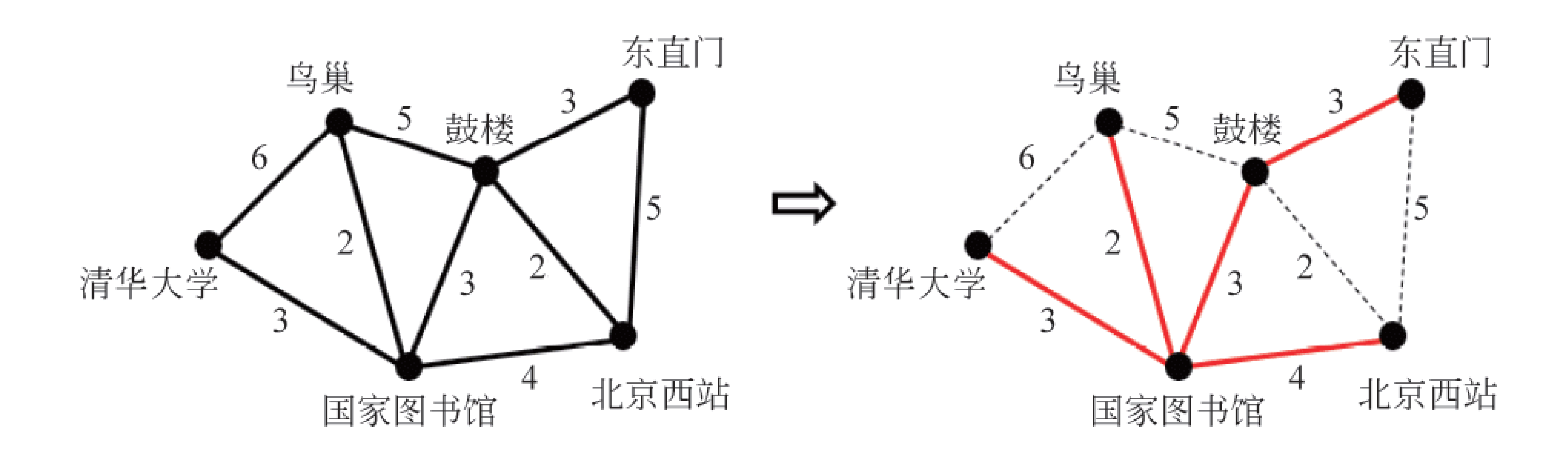


图 1.12　迪杰斯特拉算法搜索最短路径示意图

我们再用迪杰斯特拉算法来看看什么是时间复杂度。假定使用邻接矩阵来存储路径图，从算法步骤可以看到，每从一个地点出发开始搜索，都需要扫描剩下的所有地点。如果图中的节点数是 n，则需要搜索的总次数是 $n\times(n-1)$。根据前面介绍的时间复杂度的概念，可以得到迪杰斯特拉算法的时间复杂度为 $O(n^2)$。从效率上讲，随着规模 n 增大，算法效率会比较低，因此后来陆续出现了很多优化的迪杰斯特拉算法，效率有明显提升。类似地，如果采用邻接矩阵的方式存储，存储的结构是一个 $n\times n$ 矩阵，因此空间复杂度也是 $O(n^2)$。采用其他存储结构的情况，读者可以自行分析。

算法是一个很大的概念，内涵非常丰富。狭义的算法一般指具体的数学计算方法，比如加、减、乘、除，或是编写计算机程序中使用到的编程方法和技巧。广义的算法是指解决问题的具体方法和步骤，比如曹冲称象的方法。正如前文提到的，如果没有专门说明，本书中谈到的算法都是指广义算法。图灵奖得主巴特勒·兰普森（Butler Lampson）说“一切皆可计算”，通过数学建模的方式，物理现象和规律可以被计算和理解，有了互联网，人类社会也可以被计算，从 21 世纪开始，智能活动也可以被计算。当然，计算的前提是算法，算法是计算开始的基础。在解决复杂的社会经济问题时，必须从问题中抽象出可解的数学模型，围绕这个模型来设计算法，最后交给计算机进行计算，并根据计算结果修正模型和算法。在这个过程中，计算过程是由计算机负责，

更复杂的建模和算法设计,则是由人来完成。这个过程也体现了人类社会与算法之间的塑造与反塑关系。

正是因为智能经济的复杂和快速演进,人类从未像今天这样,如此渴望能够有效把握事物的未来发展趋势。幸亏当前的计算力发展日新月异,让算法的力量得以充分释放,并逐渐成为社会经济生产发展的主要推动力。

算法的基本介绍就到这里,下面我们就要依次揭开智能经济背后算法的奥秘,准备好了吗?

第 2 章　共享经济该如何共分利益

你一定听过三个和尚的故事：一个和尚挑水喝，两个和尚抬水喝，三个和尚没水喝。这个家喻户晓的故事，告诉我们合作的重要性。但是为什么两个和尚可以合作，三个和尚却不能合作了呢？其实不管两个和尚还是三个和尚，对他们来说占优策略①都是自己不挑水，等着别人挑水给自己喝。在一个共享的利益集体中，如果获得利益总是均等的，那么大多数人都会选择以最少的付出获得最大的回报。根据经济学家亚当·斯密“看不见的手”理论，在市场经济中，每一个人都从利己的目的出发，最终全社会却可以达到利他的效果。但是我们却从三个和尚的例子中，看到了“看不见的手”理论的悖论，即每个人从利己的目的出发，最后的结果损人不利己，也就是说我们得到了集体的最差解，而这个集体的最差解反而是最稳定的。为什么会这样呢？

要知道，在这个社会里，我们每个人都不会是理性的经济人，有很多因素

---

① 占优策略是一个博弈论中的常用术语，这里可以理解为在一个多人竞争或者合作的场景中，对个人最有利的应对方案。

会左右我们的行为，比如信息不对等、情感因素等。虽然我们明白合作的重要性，但是在实际生活中却不容易践行这个原则。所以这个世界每天都在发生三个和尚这样的分配问题，比如《塔木德》中“富翁的三妾争产”的故事，英法修建海底隧道工程的争端，修建机场跑道的成本该如何让各个航空公司分摊，或者无线频谱该如何分配等。虽然共享经济正在如火如荼地发展，然而不同利益主体之间的利益分配并不是一件简单的事情。面对未来的新型商业市场，共享经济的合作机制将何去何从，算法又能在其中发挥怎样的作用呢？

公平分配从来就是一道难题

财产是一切罪恶的根源；财产的分配与保卫占据了整个世界。

——列夫·托尔斯泰

人类是集体动物，人类的社会经济活动也是一种合作活动。小到一个家庭，大到一个社会，想要彼此协作创造出最大价值，首先要解决分配问题。然而，从古至今，分配都是一个不好解决的难题。

富翁的三妾争产

对于有着“世界第一商人”之称的犹太人来说，《塔木德》是他们终身都在研读的一本书，堪称犹太人的《圣经》(图 2.1)。它凝结了上千年来 2000 多名犹太学者对自己民族的历史、文化和智慧的发掘、思考和总结，是整个犹太民族生活方式的航图和精神支柱。里面不仅记载了犹太人的处世哲学，也讲了

不少理财智慧，其中有一个著名的财产分配的故事——三妾争产。

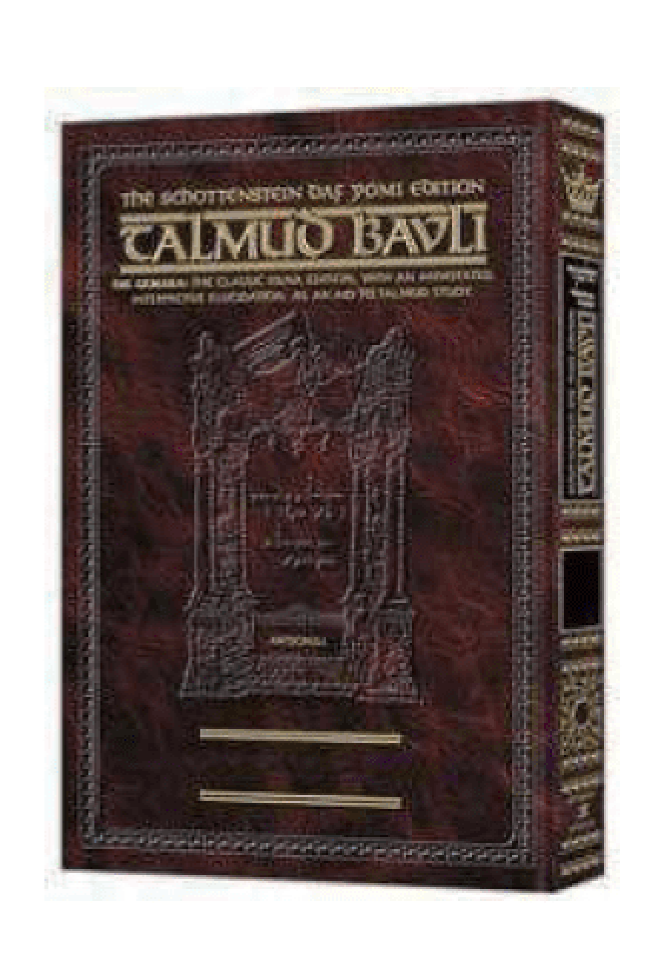


图 21 《塔木德》

一名富翁陆续娶了三房年轻漂亮的妻子，过着安逸幸福的生活。没想到年近古稀，大病一场，他这才开始为遗产的事焦虑。为了避免三个太太不必要的争端，这名富翁在遗书中对自己的财产进行了分配，他分别向这三位妻子许诺，待他死后将分给大老婆 100 块金币，二老婆 200 块金币，小老婆 300 块金币。三位太太对此也都欣然同意。

一年后，富翁病逝，然而在清算财产的时候，大家才发现这个富翁名不副实，留下来的钱根本不足 600 块金币！但是三房太太都想得到承诺的金币数量，便请来了“拉比”。拉比是一些精通律法的文士，他们会担任民事法庭的法官，进行民事案件的裁决。众拉比经过讨论，给出了如表 2.1 所示的奇怪的财产分配方法。

表 2.1　拉比们的遗产裁决

总金币数量	一　房	二　房	三　房
100 块	100/3	100/3	100/3
200 块	50	75	75
300 块	50	100	150

按通常逻辑，这三人得到的遗产比例应为 1∶2∶3，而在犹太先哲们的裁决中，只有在遗产数为 300 的情况下这一比例才成立。没有人可以解释这是为什么，这个奇怪的方案也就成了千古之谜。

英法海底隧道工程

英法海底隧道仅次于日本青函隧道，是世界第二大海底隧道，总长 50 千米，其中海底部分 39 千米，如图 2.2 所示。



图 22　英法海底隧道

这条巨型隧道的修建并非简单之事，由于成本开销和预算太大，前后经历了多次尝试与失败。

第一次尝试：1802 年，法国采矿工程师马蒂厄（Abel Mathieu）第一次提出要在英法之间修建海底隧道的这一大胆设想，标榜“2 小时马车就能互通英法”①。当时任法兰西第一共和国第一执政的正是极具军事野心的拿破仑，这个令人激动的隧道计划立刻引起了他极大的兴趣，当场拍板说要支持修建。

第一次失败（军事问题）：1803 年 5 月 16 日，由于英法两国战火不断，一场特拉法尔加战役把拿破仑的侵英梦想几乎完全击碎。这场海战的爆发使拿破仑不得不放弃了海底隧道的计划。

---

① 61 年后（1863 年），世界上才有了第一条地铁，即英国伦敦的大都会地铁，其干线长度约 6.5 km，采用蒸汽机车牵引。

第二次尝试：1834年，一个年仅27岁的年轻人，加蒙（AiméThomé de Gamond），开始对连接海峡两端的想法产生兴趣，他耗尽毕生的心血，设计了一个海底构筑带的方案，得到了当时路易王子的大力支持。

第二次失败（政治问题）：然而某天，拿破仑三世的马车在去往巴黎歌剧院的路上被炸弹袭击，经过仔细分析，得知炸弹的制作地为英国伯明翰，便再次搁置了海底隧道的推进。因为相比于一个建筑工程，毕竟还是头上的皇冠要重要得多了。

第三次尝试：从1872年开始，经过三年官僚主义斗争和外交手段的运用，方案终于在1875年8月2日得到英法两国议会的批准[①]。两国在1876年签订协议书，决定于20年期限内共同修筑海底隧道。于是，隧道从两头分别开挖，海底地质勘测等工作同时进行。

第三次失败（地理问题）：然而，施工中却遇到了一个棘手问题，1882年，鉴于莎士比亚峭壁（位于英国多佛市，是海底隧道的英国端起点）的战略地位，英国公众不愿修筑英法海底隧道，泰晤士周刊也跟着煽动民众舆论，使工程变得“妖魔化”的同时，“为了可敬的女皇和可爱的国民”，也为了不破坏“英国原有天然屏障”，这项工程在1882年7月再次被迫停止，草草收场[②]。

第四次尝试：到了1955年，英国政府宣布，海底隧道对国防安全的影响已不复存在。1957年，英法两国共同成立了英法海峡工程公司，并于1960年出具了详细报告，指出修建隧道的必要性。1964—1965年，有关研究小组再次进行地质勘探，并确定出一条隧道联络线。1966年6月8日，两国签署联合公报，第二次海底隧道建设正式开始。

---

① 英法海底隧道工程：历史过程和成功经验，李炎，牟焕森，工程研究——跨学科视野中的工程，2009(3)：90-96.

② 刘洪滨．欧洲海底隧道工程．海洋开发与管理，1990(1)：77-79.

第四次失败(成本问题)：由于实际施工过程中的造价大大超出了预算，工程遭遇到巨大的财政障碍。英法两国分别认为隧道为对方带来的好处更多，因此对方应该承担更多的成本，如此僵持，互不相让。于是在开挖 15 个月之后的 1975 年，面对无法解决的财政问题，终于有人扛不住了，英国方面最终下令停工。这样，修建英法海底隧道的尝试再次失败。

……

第 *N* 次尝试：1981 年 9 月 11 日，英国首相撒切尔和法国总统密特朗在伦敦举行首脑会晤，两国政府深刻吸取了上次的成本分担教训，决定将英法海底隧道的建设和经营承包给私人部门。这个消息一经放出，立马有人响应要接盘。1984 年 5 月，由多家银行组成的银行团向英法两国政府提交了一份关于可以完全通过私人投资来建立双孔海底铁路隧道的报告，论证了双孔两条铁路隧道的方案在技术上和财政上的可行性。银行团后来很快与英法两国的建筑公司联合在两国分别成立了 Channel Tunnel Group Limited (CTG)和 France Manche S. A(FM)公司，这两家公司再以合伙形式组成欧洲隧道公司 CTG-FM。作为一个由两国建筑公司、金融机构、运输企业、工程公司和其他专业机构联合的商业集团，CTG-FM 在 1985 年分为两个组成部分，一个是 TML(Transmanche Link)联营体，作为总承包商负责施工、安装、测试和移交运行；另一个是欧洲隧道公司(Eurotunnel)，作为业主负责运行和经营。

成功通车：解决了成本与开销的分担问题，这条隧道终于在 1994 年 5 月 6 日举办了隆重的通车典礼(图 2.3)。看来，最终拯救隧道的，不是军事战略，不是政治需求，也不是地理问题，更不是技术方案，而是其中的成本与利益分配算法！现在，每天差不多有 400 列火车经过海底隧道。截至 2013 年，通过英法海底隧道往来的列车乘客数量已经达到了 2040 万人次。



图 23　隧道通车

无线频谱的分配

锣鼓喧天，鞭炮齐鸣，红旗招展，人山人海。2016 年 10 月 21 日至 22 日，陈某某 Another Eason's Life 演唱会在国家体育场“鸟巢”开场演出，现场一片火爆，一票难求(图 2.4)。抢到票的粉丝难掩内心的激动，此刻最重要的事恐怕就是发个朋友圈。

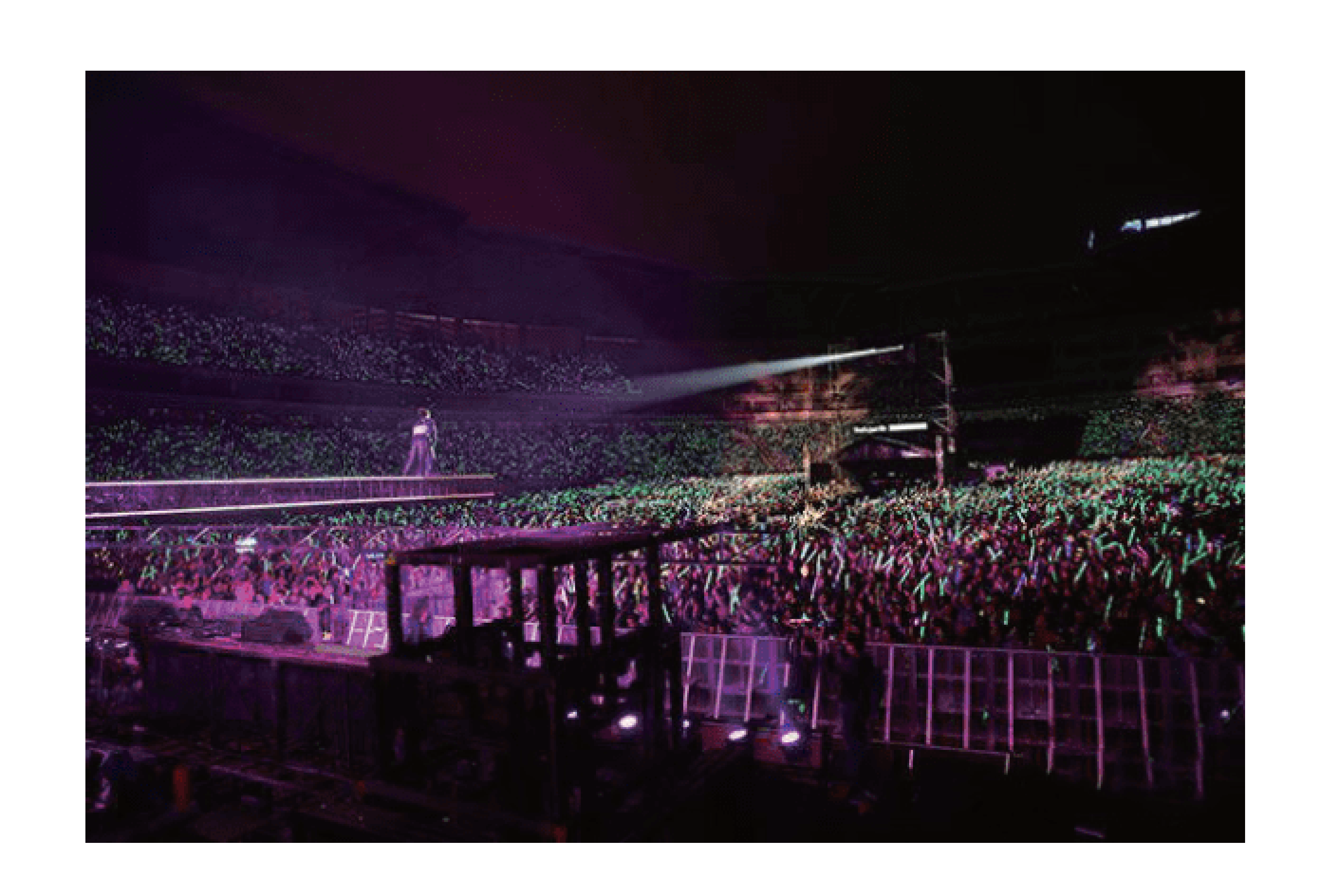


图 24　演唱会现场

然而别说是上网了，这时连个电话都很难拨出去……当你欲哭无泪的时

候，请环顾四周，看是不是有类似图 2.5(a)的这种车？一句话：靠近它！这可能是你唯一能在朋友圈展示与偶像近距离接触的机会了。因为它就是传说中的移动基站(信号车)。通常在人流量剧增的时候，为避免通信拥塞，主办方都会与运营商协调，派出信号车增援，开通临时信道。



(a)



(b)

图 25　无线信号车与基站树

信号车到底是怎么给我们提供网络接入的呢？首先我们知道提供无线覆盖的是基站。通俗地讲，基站就是公开场合的大路由器，用来实现有线通信网络与无线终端之间的无线信号传输。

通常情况下会是图 2.5(b)这样，是一棵伪装成树的基站，上面密密麻麻的“叶子”都是采用 ABS 塑料或 PU 等高性能防阻燃环保材料制成，不仅环保，而且防紫外线。基站在通信网络中的位置如图 2.6 所示。看不太懂？没关系，你只需要知道，基站负责把你的手机接入运营商网络，之后的事情就是运营商来全权处理啦。

大家之所以能够在家里、学校、地铁、大街、商店里随时上网，都是因为有密集的基站网络。然而，随手一搜，周边就能有十几个 Wi-Fi，它们无时无刻不在收发着各种无线信号，如此多的基站为什么能够互不干扰？我们收到的消息为什么不会错乱？这就要感谢一位漂亮得不像实

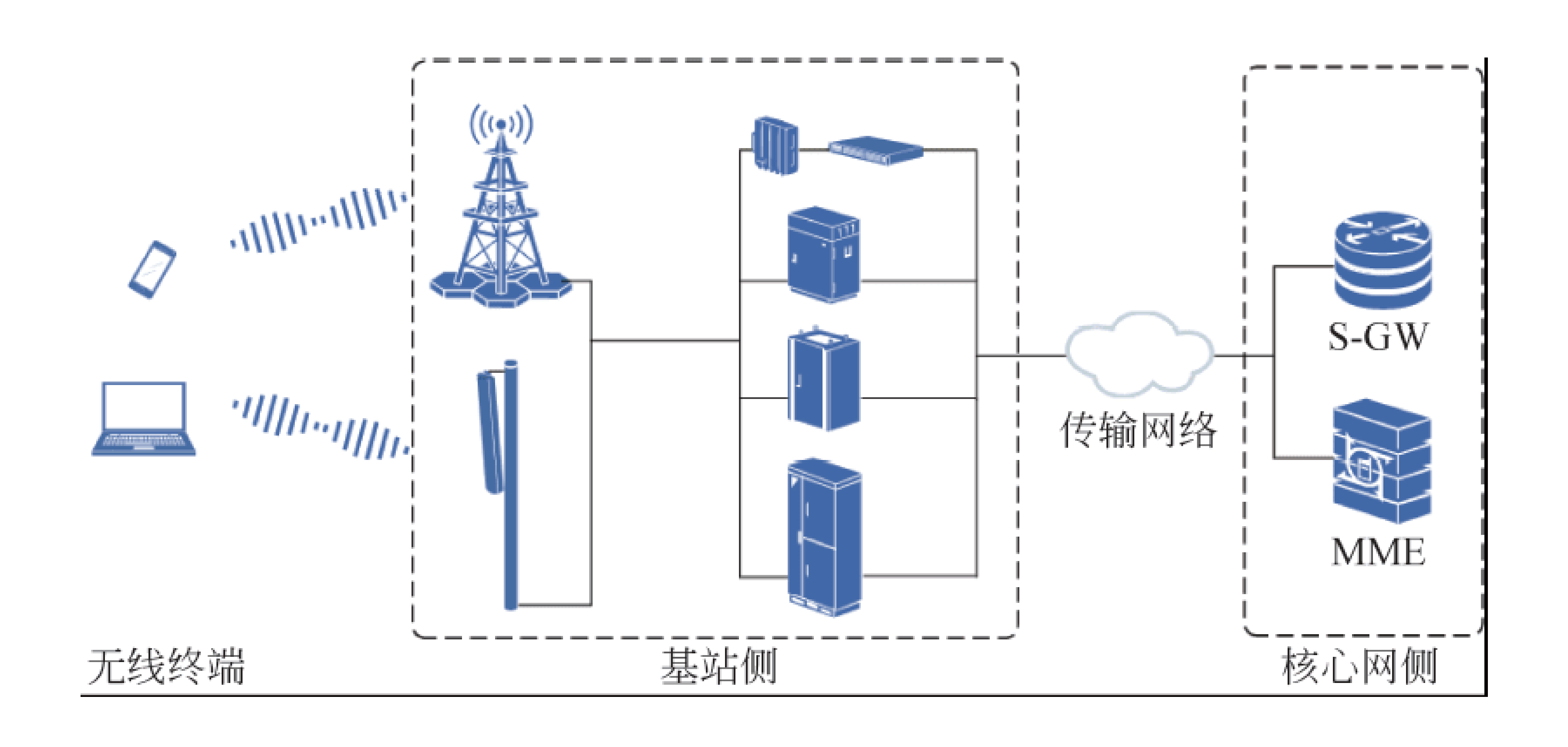


图 26　基站在通信网络中所处的环节

力派的女神。

她曾是艳绝一时的明星，被称为“世界上最美丽的女人”。她又是一位女发明家，她与他人合作发明了“扩频通信技术”，被广泛用于今天的手机、卫星通信和无线互联网。2003 年，波音公司做了一系列的宣传广告纪念这位传奇的好莱坞女明星、科技女性。她就是海蒂·拉玛(Hedy Lamarr)，一位美国影视女演员，跳频技术的第一发明者，首次将有线转为无线，被后世尊为“CDMA 之母”。2014 年，海蒂·拉玛入选美国发明家名人堂(National Inventors Hall of Fame)，图 2.7 是海蒂·拉玛和她发明的跳频技术专利①。

但是，可以在无线通信中使用的频段只是电磁波频段中很小的一部分，因为有很大部分的频段是不适合提供数据通信的。因此，这宝贵的无线通信频段历来为商家必争，谁获得的频段宽，谁就能用更优质的服务提供人们随处上网，也就能收取更多的流量费。那么，到底该如何在不同运营商之间分配无线通信频段呢？很多国家都做过各式各样的探索，其中最成功的当属英

① 细心的读者可能注意到，发明专利上的名字是 Hedy Kiesler Markey，没错，这就是海蒂·拉玛，她当时处于已婚状态，这是她当时随丈夫姓改的名字。

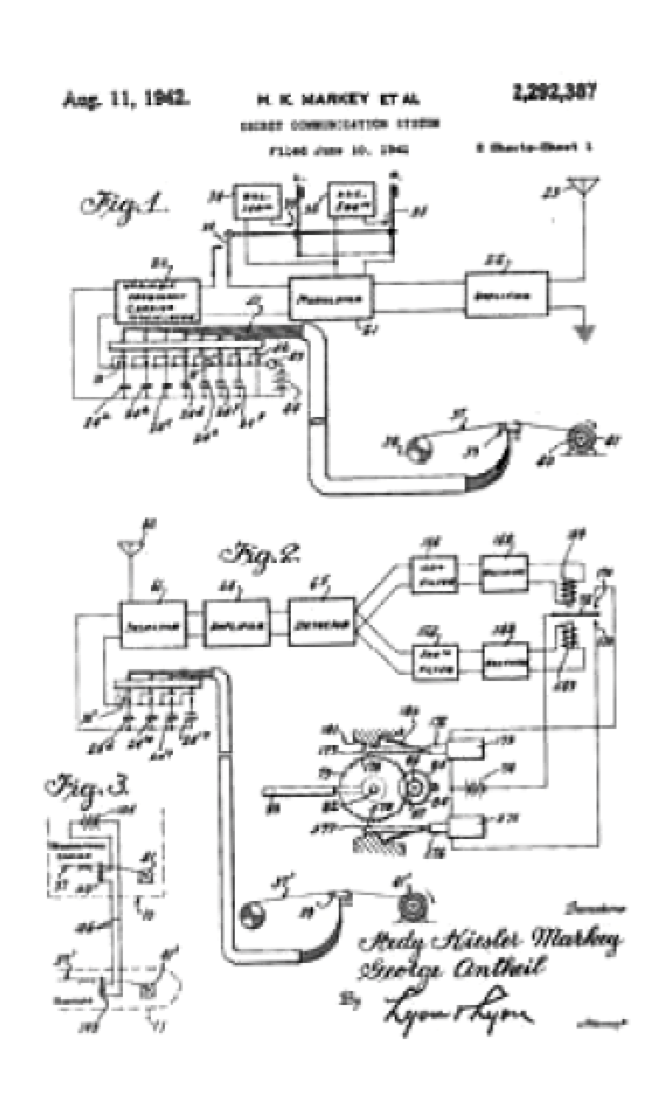



图 27　海蒂·拉玛和她发明的跳频技术专利（来自美国发明家名人堂）

国的 3G 频段拍卖，获得了 225 亿英镑的收入，占当年英国 GDP 的 2.5%。当然，政府的高额收益，就意味着是运营商的惨痛出血，这场在政府圈子被树立为榜样的 3G 频谱拍卖最终导致了英国的电信泡沫。这场拍卖也成了一个不可复制的神话。

下面就让我们来看看神话是如何产生的。2000 年，英国政府组织了一次有史以来最大的 3G 频谱拍卖，将 4 个 3G 频段进行拍卖。当时有 4 个已经具备很强竞争力的 2G 运营商，享有非常大的价格优势。因此，政府考虑采用升价拍卖，这样可以阻止其他公司进入后阶段的拍卖。这次拍卖采用了升价拍卖中的“英式拍卖”和“荷兰式拍卖”的综合模式，随着升价拍卖的进行，频谱价格在不断提高，很多小公司由于成本不足陆续退出竞争，在只剩下 5 家公司时不再采用升价拍卖，改换成“封闭式拍卖”。在这种拍卖方式下，每家公司再也看不到其他人的出价。由于每段频谱只能卖给一家公司，而且每家公司也只能购买一段频谱，为了避免成为最后一名而被淘汰出局，大家都不得不按照自己的能力极限来出价。就这样，英政府通过非常有利的拍卖

规则，看到了5家公司的底牌，卖掉的这4段频谱堪称价值连城。

英国的巨额收益极大地刺激了其他国家，它们在看到英国政府从3G频谱中获得了巨大成功以后，争相效仿。在随后的几年里，不断有各个国家打算靠出售频谱来提高GDP，而且多半都在效仿英国的升价拍卖。

拥有5个运营商的荷兰紧接着就宣布要出售5个频段，既然运营商和频段数量相等，那么他们也决定采取升价拍卖的方式，并预期那些边缘化的小运营商一定会在升价拍卖的过程当中逐步退出竞拍，从而最终在大公司身上大赚一笔。

在意识到自己的危险处境之后，一些很有潜力的边缘公司开始与在位官员私下沟通，而荷兰的竞争政策和拍卖设计得不完善，导致一些其他非荷兰本土的企业也来与当地官员搞合作（例如Deutsche Telekom、DoCoMo等等）。结果这场竞拍最终只收益了30亿欧元，与预想的100亿欧元相差甚远。

随后，2000年10月份的意大利，11月份的瑞士，2001年的比利时、希腊……大家纷纷来试水，但均以失败告终，都没能再次触及英国政府的价格高峰。

什么样的拍卖才能获得成功呢？纵观所有上述拍卖，为什么即使相同的策略也会产生完全不同的竞拍结果？参与竞拍的运营商们难道都是针锋相对、你死我活吗？有没有什么样的频谱分配方法能够使运营商们共赢，又同时能创造最大的社会效益，从而皆大欢喜呢？

谁来出钱修跑道

香港在回归之前原本有个老机场，名叫启德机场。启德机场位于九龙半岛，跑道是在香港岛和九龙半岛之间的维多利亚海湾里填海造成的。香港岛

和九龙半岛都是摩天大楼林立，飞机起降于港岛和九龙半岛的两大群水泥森林中间的狭长空隙走廊，20世纪90年代的港片中经常出现的飞机起降镜头，都是在这里拍摄的。随着香港经济的发展，启德机场相当繁忙，客货载运量已趋饱和，由于受地理位置的限制，已没有扩展的空间。

1989年10月，港英当局忽然单方面提出了一个香港有史以来最庞大的基础设施建设计划——《玫瑰园计划》，其中包括在港岛西南填海兴建新机场，建设与新机场配套的通往港岛的机场铁路和高速公路、桥梁以及港岛西区海底隧道等，总投资约上千亿港元，预计十多年才能完工。

这些计划公布后，香港和外来的投资者都不敢贸然参与，因为这些工程跨越1997年6月30日香港回归中国之后，涉及未来的中华人民共和国香港特别行政区政府（以下简称香港特区政府），而没有中国政府的明确支持，这些工程的融资就困难重重。为了解决这一难题，英方才不得不找中方商谈。

中方从香港的现实需要和长远发展考虑，赞成兴建香港新机场（建成后的新机场如图2.8）。中方主要关注的是港英当局不能在香港回归前把财政储备都花光，而应给未来的香港特区政府留下足够的财政储备。建设新机场应讲成本效益，不应让未来的香港特区政府背上沉重的债务包袱。



图28　从大屿山俯瞰的香港国际机场

1990 年 7 月，英国外交国务大臣弗朗西斯·莫德访华时，中英双方同意成立专家组共同研究修建香港新机场问题。同年 10 月，中英双方正式开始谈判①。然而，1992 年 4 月，英国派出彭定康取代卫奕信出任香港总督。彭定康在到任半年之后的 1992 年 10 月 7 日，公布了香港 1994 年的区域组织选举和 1995 年立法局选举的新方案，把立法局的间接选举改为变相直接选举，破坏了“直通车方案”。然后，他才访问北京，与国务院港澳事务办公室主任鲁平会晤，结果会晤中没有取得任何成果。随后，鲁平召开记者招待会，批评彭定康的方案“三违反”：违反《中英联合声明》，违反《基本法》，违反两国外交部长和外交大臣达成的协议，“要成为香港的千古罪人”。

总之，在新机场建设与成本分摊的问题上，中英双方一直存在着激烈的斗争。

让我们假设有 5 家航空公司要准备联合建造一个新机场，但这几家航空公司飞机的机型不同，对跑道长度和承受压力等要求也都各不相同，这无疑会增加成本分摊的复杂度。

由于适合大型飞机的跑道必定可以容纳小型的飞机，然而反之不然，为方便起见，我们假设每家航空公司只拥有单一机型，各家公司的机型互不相同。按照机型由小到大，我们用 A、B、C、D、E 来描述这 5 家航空公司，很显然，A 公司单独建立一个小型跑道，足以供自己小型飞机的起降，假设其成本为c_A。而如果 E 公司想要单独建设一个跑道，要花费的成本c_E 远远大于c_A。常识告诉我们，这 5 家公司如果共同修建这个大跑道的话，总成本仍然是c_E，所以在一般情况下，共同修建跑道将使得每家公司都有可能节省一笔成本费用。

但是问题来了，这 5 家航空公司到底该如何分摊这c_E 的总成本呢？如果通过谈判和协商，能够达成合作协议，那么便可以皆大欢喜。然而，倘若谈判

① 姜恩柱. 大国较量：中欧关系与香港回归亲历. 北京：中信出版社，2016.

破裂，事情也未必会演变成 5 个大大小小不同的跑道，因为这 5 家航空公司必然会形成子联盟。设想一下，若 A 公司和 E 公司联手合作建成一个机场，它们采用了一个非常简单的成本分摊，分别承担$c_A/2$ 和$(c_E-c_A/2)$，那么这个结果已经明显低于各自独立建设机场跑道的成本 c_A 和 c_E。这样一来，在这场 5 家航空公司的谈判过程中，A 公司和 E 公司的“底线”就已经降低到了 $c_A/2$ 和$(c_E-c_A/2)$。然而，类似这样的子联盟可以有许许多多，参与公司的数量又不限于两家，因此，我们不得不面临众多联盟之间的相互作用问题，因为这些子联盟之间的博弈关系，决定着每家公司在成本分担中能够接受的价格“底线”，也就决定着能否达成最终合作建设跑道的协议。

经过中英联合联络小组和专家组等一系列会谈，双方于 1991 年夏达成协议，形成了一个书面文件，即《关于香港新机场建设及有关问题的谅解备忘录》。该谅解备忘录确定，新机场建设要符合成本效益，本着不在财政上给香港特区政府造成负担的原则。具体规定：在 1997 年 6 月 30 日前，港英政府“将在最大程度上完成”机场核心项目建设；跨越 1997 年 6 月 30 日偿还的债务总额不超过 50 亿港元，如超过，须由双方磋商；港英政府将预留 250 亿港元财政储备给未来的特区政府①。

功利主义的分配方案——Shapley 值

一个人的价值，应当看他贡献什么，而不应当看他取得什么。

——爱因斯坦

① 姜恩柱. 大国较量：中欧关系与香港回归亲历. 北京：中信出版社，2016.

通过上面几个例子，我们已经看到，争端以及争端能够得以解决，合作以及如何合作，都离不开一个前提，那就是参与各方的利益分配问题。这些问题同样出现在今天的共享经济里，但幸运的是，今天我们可以用 Shapley 值来精确计算出参与方做出的贡献以及该如何更加公平地分配合作收益。

Shapley 值，一个天才提出的天才理论

提出 Shapley 值的罗伊德·沙普利（Lloyd Shapley，下文简称沙普利）是位传奇人物。他出生于 1923 年 6 月 2 日的美国麻省剑桥镇，一个有着浓郁科学氛围的家庭。沙普利的父亲哈罗·沙普利（Harlow Shapley），是一位著名的天文学家。作为哈罗·沙普利五个孩子中的第四个，沙普利在哈佛大学天文台台长官邸长大成人，很早就显露出数学天分。沙普利后来在回忆中提到：

“我有两个聪明的哥哥，他们都是课程全优的学生，并且我猜我的姐姐也很聪明。然而，有时候我们在家里会用纸牌玩诸如数学乘法运算有关的游戏，我能够从想方设法胜出他们的过程中得到一种激励。虽然他们比我大四到六岁，但我的表现相当出色，由此我在家里得到一个数学奇才的名声。”

1943 年，在读大二的沙普利被送到位于北卡罗莱纳的新兵训练营，开始了三年军旅生涯。在军队服役期间，沙普利会写信回家，当然信中的一切内容都会受到审查，他甚至不能让家人知道他在什么地方。为了让家人们知道他的位置，他构思了一个真实的探测信件，在其中提到一些事情如他的“查理叔叔（Uncle Charlie）”，这些事情不会受到审查员的注意，但是会向家人传达出信息，这封信的含义并不在单词里（家里从来就没有过一个名叫“查理”的叔叔）。后来他的哥哥威利士·沙普利（Willis Shapley）想到，原来它的意思在于信中每一行的第一个字母：C-H-I-N-A，从而知道沙普利那时已经到达

了遥远的中国。

服役结束后，沙普利回到了哈佛大学，他曾一度迷茫，直到他被兰德公司录用。

兰德公司在公众中的知名度并不高，但它的实际影响力却很大。兰德公司被称为美国的“智库”，它影响和左右着美国的政治、经济、军事、外交等一系列重大事务的决策，有意思的是，它是一个非营利的民办研究机构，独立地开展工作，与美国政府只是客户合同关系。兰德公司的研究成果举世瞩目。已发表研究报告 18 000 多篇，在期刊上发表论文 3100 篇，出版了近 200 部书。在每年的几百篇研究报告中，5％是机密的，95％是公开的，而这 5％的保密报告随着时间的推移也在不断解密。这些报告中，包括“中国 21 世纪的空军”“中国的汽车工业”“日本的防御计划”“日本的高科技”“俄罗斯的核力量”“韩国与朝鲜”“数字化战场上的美国快速反应部队”等重要课题。

“军方认为应该在战后继续与科学家群体保持联络，给他们分配一些任务和资金，让科学家思考问题并告诉军方有关的一切。这是一种广泛而开放的与空军之间的合同。这导致一些人如约翰·廉斯把相当多的各种各样的人组织在一起。他从数学系雇用了一些疯狂的天才学生，包括我。”沙普利在接受采访时对记者说。

在兰德公司(美国政府的智囊团公司)中有一个小组决定研究博弈论。他们组织研讨会，每周见面，研读《博弈理论和经济行为》其中的一章。这本书最初在 1944 年出版，标志着博弈论这一新的数学分支的正式创立，从此，当我们需要分析多个有利益冲突的参与者之间的合作与竞争问题时，就有了新的工具。

“冯·诺依曼和奥斯卡·摩根斯坦恩这本书的出版，是相当了不起的事情。我采用 1947 年的版本即原书的第二版，其中只是简单地增加了一些附录。这本书问世以来，没有引起太大轰动和广泛的评论，无论如何，冯·诺依曼已经是大名鼎鼎的科学家，但是除了这些以外，并没有什么事情发生。”

由于参加研讨会，沙普利得以与另外一位数学家埃尔文·罗斯（Alvin E. Roth）一起工作，解决该书中提出的一个问题，怎样寻找矩阵博弈所有的解。他们共同解决了这个问题，因此它被命名为 Shapley-snow 解。然而，这是沙普利在没有阅读大量数学论文情况下完成的，他对这个解能够用来做什么也没有特别清晰的概念。

冯·诺依曼（Von Neumann）读了他们的文章，非常感兴趣，因为在那个时期很少有论文讨论有关博弈论的问题。

那时冯·诺依曼或许出于个人自尊，尽管他已经不再做有关博弈论的工作，还是对他们的工作鼓励一番。他那时主要在思考计算机方面的问题，他想鼓励他们继续去做，为此写了热情洋溢的评论，表示他对此非常兴奋，他将会出版它，也将会推荐一个指定的期刊去发表等等。

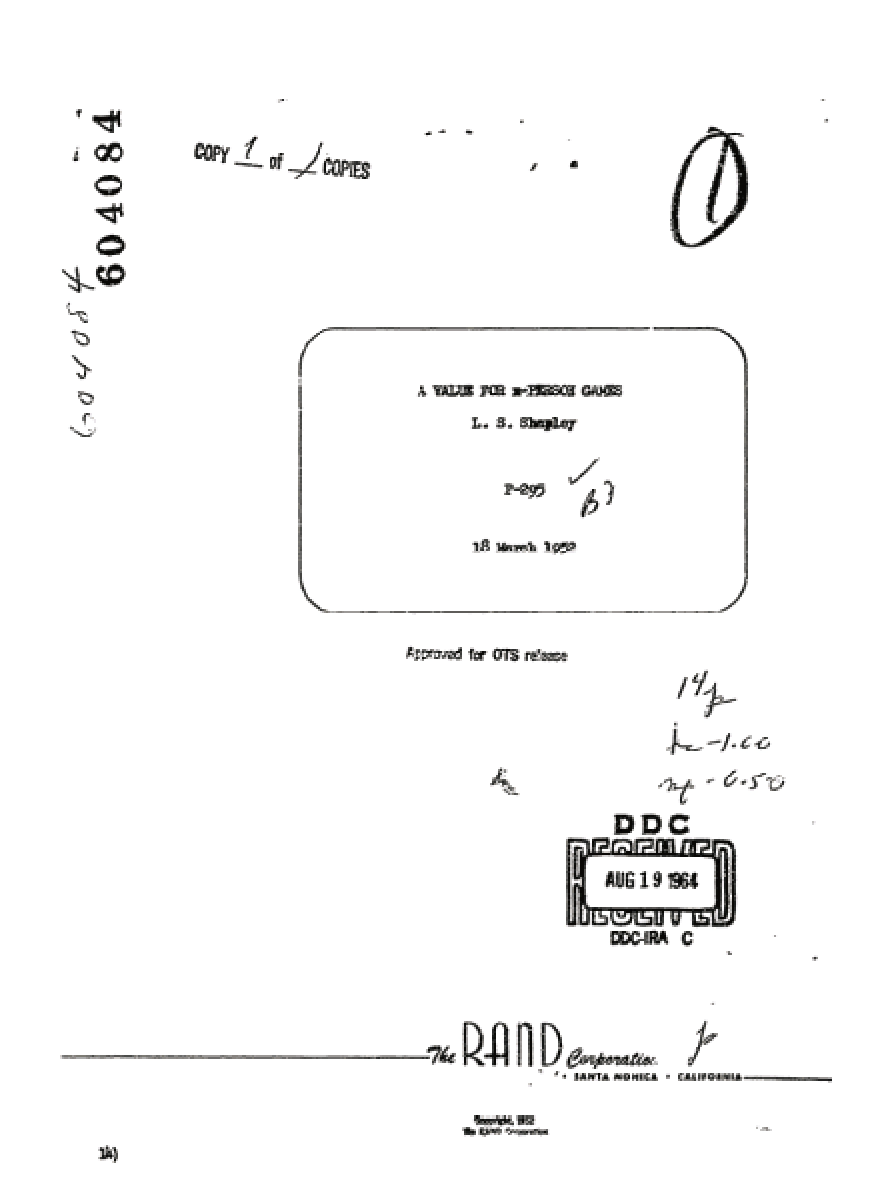
在兰德公司工作不到两年后，沙普利前往普林斯顿大学继续研究生学习，在数学家艾伯特·图克（Albert Tucker）的指导下沙普利完成了博士论文"集合函数的可加性和非可加性"。同一时期在那里他发表了几篇其他方面的论文，包括 1953 年发表的一篇名叫"多人博弈的价值（*A Value for n-person Games*）"的文章，正是这篇文章引入了如今被大家广泛知晓的"Shapley 值"，这是一个有关合作博弈解的概念。

沙普利在创立 Shapley 值这个概念后不久便开始考虑它的应用。他与马丁·苏比克（Martin Shubik）一同工作，将它用于在选举中度量影响力。这些工作导致了一个被称为 Shapley-Shubik 影响力指数的出现。他们两人作为不知名的研究生，一个学习数学，一个学习经济学，有点不知天高地厚地将这篇论文提交给顶级的政治科学杂志，令人吃惊的是在几周之内就被接受了。

图 2.9 是沙普利在 2012 年诺贝尔经济学奖颁奖典礼上做演讲的照片和他 1953 年发表的 *A Value for n-person Games* 文章封面。请注意，沙普利获得诺贝尔奖的贡献是他在稳定匹配方面的研究成果（第 3 章会介绍），但作者

认为 Shapley 值至少和他在稳定匹配方面的成果同样重要，如果不是更重要的话。




604084

COPY 1 of 1 COPIES

A VALUE FOR n-PERSON GAMES

L. S. Shapley

P-295

18 March 1952

Approved for OTS release

DDC
RECEIVED
AUG 19 1964
DDC-IRA C

The RAND Corporation
SANTA MONICA · CALIFORNIA

图 2.9　沙普利在 2012 年诺贝尔经济学奖颁奖典礼上的照片和他 1953 年发表的 *A Value for n-person Games*

用边际贡献率解决分配的公平性问题

一项工作，只要有多个参与方，那么就涉及合作博弈和利益分配问题，Shapley 值的最大贡献就是解决了合作博弈中各方的利益分配。

我们用一家农产品供应商来举例。如果他准备卖掉刚刚收获下来的苹果，他可以自己去市场上卖，也可以拿到农产品展销会上去卖。如果他自己去市场上卖的话，这些苹果可以卖 2000 元钱；如果拿到展销会上去卖，可以卖到 4000 元钱。对于农产品展销会来说，如果不接受这家供应商的苹果，而是接受其他小商家的农产品，收入仅为 1000 元。这样一来，如果两家达成合作，展销会上售卖这家供应商的苹果，那么两家总收益为 4000 元；如果各自

为营，总体收益为3000元。显然，从总体收益来看，供应商和展销会合作才是最有效的。那么，在合作过程中，苹果出售赚取的4000元应该如何分配？供应商和展销会应各拿多少？

为解决这个问题，我们首先要了解双方的谈判力度如何，也就是说，双方对于达成合作做出的贡献大小，博弈论中将这样的贡献称为**边际贡献**。某个参与人的边际贡献就是指他参与合作与不参与合作，对整体效益产生的利润之差。在这个例子当中，缺少供应商或者展销会任何一方的合作，总体的收益之和都将变为3000元，所以双方的边际贡献都是1000元，也就是说，离开任何一方，这1000元的增值都不会实现。这表明，在这场合作当中，双方的边际贡献一共是2000元，各自占据1/2，双方的谈判能力是对等的。那么，当两个人的边际贡献相等时，合作带来的剩余收益应该平均分配。因此，在这4000元的利润分配中，供应商应得到2500元，展销会应得到1500元。

那么，当有更多的参与方想要加入，各方组成大联盟的时候（例如，有广告商想趁此展销会的机会做产品宣传），仍然需要计算各参与方的边际贡献，那么就需要将所有参与人进行排列，然后让他们按照排列顺序逐个加入联盟，依次计算当他加入到这个联盟之时，为联盟做了多少边际贡献。但问题是，排在后面进入联盟的人往往会占很大优势，因为每个人的边际贡献计算的是当他进入联盟时，他与前面所有已经进入联盟的人联盟所产生的边际获利。因此，每个参与人都希望自己加入联盟的位置尽量靠后。

为了解决这个问题，沙普利设想了一个办法，可以使每个人都处在一个有相等可能性的顺序位置上。具体做法是，将每位参与方的Shapley值定义为在全部$n!$种可能的排列顺序下，每人对应的边际盈利向量的算术平均值。在排列σ中，如果用$P^{\sigma}(i)$表示排在参与方i前面进入联盟的人的集合，$v(S)$表示联盟S的总体收益，那么参与人i在这个n人联盟的Shapley值ϕ_i可以简单地表示为：

$$\phi_i = \frac{1}{n!}\sum_{\sigma\in\pi(N)}(v(P^\sigma(i)\cup\{i\})-v(P^\sigma(i)))$$

其中，$v(P^\sigma(i)\cup\{i\})-v(P^\sigma(i))$表示在参与人 i 加入联盟之时，产生的边际效用。对于 $n!$ 种排列，取边际效用的平均值，即为参与人 i 在该联盟中的 Shapley 值。

公式太复杂没看懂？没关系，看下面的例子，你很快就会明白。

程序员鼓励师

Shapley 值的用途十分广泛，实际上，凡是多人合作需要商量收益如何分配的场景，都可以尝试使用 Shapley 值。

程序员鼓励师一直被认为是一个网络段子，但近年来却成为越来越多互联网公司造福员工的常用手段。“男女搭配，干活不累”，有业内人士指出，公司适当增加女员工比例，确实有助于办公室氛围的和谐。

知乎网上一位网友给出了如下一个问题：

某互联网公司今天加班，需要编写一个 500 行的程序代码，产品经理找了三个程序员来完成，按照完成量发奖金：1 号普通程序员独立能写 100 行，2 号大神程序员独立能写 125 行，3 号美女程序员能写 50 行。但如果程序员两两合作，会产生不同的编码效率：1 与 2 号合作能写 270 行，2 与 3 号合作能写 350 行，1 与 3 号合作能写 375 行（请自行脑补：美女都是催化剂，码农都是潜力股）。当然，三名程序员共同合作能完成 500 行。

若共有 1000 元项目奖金，该如何给这三名程序员分配呢？

下面，我们尝试用 Shapley 值进行计算。首先，计算可能的联盟数量。显然，三个人的联盟形成方法一共有 6 种：

（1） 1 号邀请 2 号加入组成 S 联盟，3 号加入 S 联盟；

（2）1 号邀请 3 号加入组成 S 联盟，2 号加入 S 联盟；

（3）2 号邀请 1 号加入组成 S 联盟，3 号加入 S 联盟；

（4）2 号邀请 3 号加入组成 S 联盟，1 号加入 S 联盟；

（5）3 号邀请 1 号加入组成 S 联盟，2 号加入 S 联盟；

（6）3 号邀请 2 号加入组成 S 联盟，1 号加入 S 联盟。

按照 Shapley 值的计算过程，下一步需要计算每位程序员的边际贡献，如表 2.2 所示。

表 2.2　程序员边际贡献

可能性	加入顺序	1 号的边际贡献	2 号的边际贡献	3 号的边际贡献
1/6	1—2—3	$v(\{1\})=100$	$v(\{1,2\})-v(\{1\})=270-100=170$	$v(\{1,2,3\})-v(\{1,2\})=500-270=230$
1/6	1—3—2	$v(\{1\})=100$	$v(\{1,2,3\})-v(\{1,3\})=500-375=125$	$v(\{1,3\})-v(\{1\})=375-100=275$
1/6	2—1—3	$v(\{1,2\})-v(\{2\})=270-125=145$	$v(\{2\})=125$	$v(\{1,2,3\})-v(\{1,2\})=500-270=230$
1/6	2—3—1	$v(\{1,2,3\})-v(\{2,3\})=500-350=150$	$v(\{2\})=125$	$v(\{2,3\})-v(\{2\})=350-125=225$
1/6	3—1—2	$v(\{1,3\})-v(\{3\})=375-50=325$	$v(\{1,2,3\})-v(\{1,3\})=500-375=125$	$v(\{3\})=50$
1/6	3—2—1	$v(\{1,2,3\})-v(\{2,3\})=500-350=150$	$v(\{2,3\})-v(\{3\})=350-50=300$	$v(\{3\})=50$

由表 2.2 可知：

1 号普通程序员的 Shapley 值为：

$$\frac{1}{6}(100+100+145+150+325+150)=\frac{970}{6}$$

2 号大神程序员的 Shapley 值为：

$$\frac{1}{6}(170+125+125+125+125+300)=\frac{970}{6}$$

3 号美女程序员的 Shapley 值为：

$$\frac{1}{6}(230+275+230+225+50+50)=\frac{1060}{6}$$

三人 Shapley 值的总和正好等于 500。

所以，根据 Shapley 值，1 号普通程序员应该获得的奖金为：1000×32.33％＝323.3 元，2 号大神程序员应该获得的奖金同样为 323.3 元，3 号美女程序员应该获得的奖金为总奖金的 35.33％，即 353.3 元。看来，在这个算法定义的世界，长得美真的是可以当饭吃啊！

我们曾在 2.1 节提出了无线频谱分配的问题，读者同样可以试着用 Shapley 值计算一下，或者借鉴文献 *Performance and Incentive of Teamwork-based Channel Allocation in Spectrum Access Networks*（IWQoS'2015）的解法。在我们的赛博新经济公众号的推送中也可以寻找到答案。

平均主义的分配方案——核

财富分配的差异与不均是最普遍、最持久的冲突之源。

——詹姆斯·麦迪逊，作家

什么是核

Shapley 值是将收益按照参与人的边际贡献率进行分摊，参与人应获得的收益等于该参与人对每一个他所参与联盟的边际贡献的平均值。那么换一个角度，如果各方贡献相等时，又该如何分配？

让我们回顾一下 2.1 节中机场跑道成本分摊的例子。对于 5 家航空公

司 A、B、C、D、E 的集合 N，不论以怎样的子联盟来划分（S_1，…，S_k），每家航空公司都能感觉到大联盟 N 中存在一个分配，这个分配赋予他的盈利至少不差于他所属联盟 S_i 能分配给他的盈利（在这个例子当中，分担的成本为“负盈利”，即要求负盈利越少越好），那么我们就可以称大联盟或者联盟博弈具有内聚力。

对于具有内聚力的联盟博弈，如果一个配置 x 满足如下条件：不存在任何 S 属于 2^N，使得这个 S 可以改善 x，那么满足这个条件的配置全体就是原问题可行配置集的一个子集，我们称这些配置为这个博弈问题的一个解，这就是“核”。

“核”是沙普利提出的又一个解决分配问题的理论。是的，沙普利的一生就是一个大写的天才成长之路，他除了提出了被人熟知的“Shapley 值”之外，还提出了令他获得诺贝尔奖的 Gale-Shapley 匹配理论（第 3 章将会提到）。然而，他还有一个更卓越的成就，就是他提出了“核”的概念。“核”从平均主义的角度来衡量各个不同联盟所能带来的福利，衡量的标准为联盟的剩余利润。Shapley 值是将收益按照参与人的边际贡献率进行分摊，参与人所应当获得的收益等于该参与人对每一个他所参与的联盟的边际贡献的平均值。Shapley 值强调边际贡献，核则更偏向于公平。Shapley 值体现的是一种功利主义的“公平”，而核则体现的是平均主义的公平性，这是一种同情与保护弱势群体的分配方案。

沙普利曾经回忆说：“兰德公司从不试着将我拉入与战争博弈有关的事情或此类问题的研究，对我来说，兰德公司最了不起的地方是他们让我做我想做的事情，我能成功地做我想做的。最终我得到了来自美国国家科学基金会的资助（National Science Foundation，NSF），我在兰德公司工作，但并没有从兰德公司领取薪水。我在兰德公司比我应该待的时间多了大约 10 年，因为在那段时间，我的所有支持来自 NSF，这可以让我做任何吸引我的研究。”

“核”来帮你卖鞋

有三个局中人甲、乙、丙，甲有一只左鞋，乙和丙各有一只右鞋，这只左鞋和任一只右鞋能相互匹配成一双鞋，售价 100 元。单只鞋或两只右鞋由于无法使用而一文不值，所以甲、乙、丙三人所拥有的三只鞋总价值仍然为 100 元。那么在这场售卖当中，三人应该如何分配这 100 元的盈利呢？

很显然，甲在这场博弈中显得尤为重要，因为如果没有甲的参与，那么乙和丙即便组成联盟，他们拥有的鞋的最终价值仍然为零。

“核”概念为这个联盟博弈给出了一个唯一的解，结果令人惊讶：甲、乙、丙三人的利润分配为(100，0，0)。在这个结果中，甲完全不会受到乙和丙联盟的威胁，因为他们的联盟并不能改善自己的被动境遇。对于乙和丙来说，只要他们是理性的，如果甲对其中任何一个人给予哪怕是一丁点的鼓励，例如，甲同意支付给乙 1 元钱作为回报，于是分配结果变为(99，1，0)，便足以“引诱”乙与甲组成联盟，从而破坏掉与丙组成的零收益联盟。但是这种方案仍然是不稳定的，因为甲又再次能和丙组成新的联盟，使得上述分配方案得到改善，因为乙背叛了与丙的联盟，丙表示只需要 0.5 元的收益分配就同意与甲组成联盟，因此甲便被丙引诱而组成了新的分配方案(99.5，0，0.5)，这个分配方案使得甲和丙的处境都得到了改善。如此递推的话，乙和丙会一直以更小的收益期望引诱甲与自己组成联盟，因此最终必将收敛到方案(100，0，0)，才能使得甲不再会寻求改善方案，而乙和丙的联盟又不会对这个方案形成任何威胁。此时，甲得到 100 元，乙和丙分别收获 0 元，便构成了这个博弈唯一的解。

直观上看起来，这个解似乎有点荒谬，但更为荒谬的是，如果有 100 人拥有左鞋，101 人拥有右鞋，比起这个庞大的基数，多出来的 1 只右鞋似乎是微

不足道的，人们直观上可能会认为，此时的左鞋与右鞋应该值差不多等值的钱。但核的概念仍然会给出唯一的一个解：若某参与人拥有左鞋，获利 100 元，若某参与人拥有右鞋，获利 0 元。

这种“逆直观”的分配方案与只有三个参与人时的利润配置竟然一样，虽然有悖于直观，但它基本符合经济学的解释：右鞋的供应一旦超过左鞋的供应，那么右鞋便将一文不值。这个配置方案的证明也非常简单。这 100 双鞋的总价值为 10 000 元，假设有一个拥有右鞋的人分得的利润为 1 元，那么剩下的这 200 个参与人便只能分得一共 9999 元，此时，这 200 个参与人不如排挤掉这个获利 1 元的右鞋持有人，在他们这个 200 人的子联盟中共同分配这 10 000 元的利润，那么这样就能多获得 1 元的收益，从而改善这 200 人的处境。与上面三个参与人的推理相同，这样的改善会一直持续下去，因此，在最后属于“核”的分配方案中，右鞋的持有人必将获得零收益。

由于这个结果与人们的直观想象离得比较远，有一些人因为这个例子对“核”提出了责难，但笔者认为，“核”的概念还是很吸引人的，它毕竟满足人人都认可的理性条件，它最大的缺陷应该在于解数量的不确定性，因为有时候核可能是空集，也可能有唯一解，还可能包含许许多多个解，甚至无限多个解。在这一点上，Shapley 值作为一个“一点解”，受到了更多研究人员和应用人员的青睐。实际上，这两个概念是从两个不同的角度分别解决公平分配的问题。

三妾争产——千年难题是如何得到解决的

下面让我们回到前面提到的三妾争产问题。直到 1985 年，才有人解决了这个问题，算是弄明白了一千多年前的“拉比”们的解决思路。

设有三位继承者，所有继承者要求的财产从少到多分别记作 $c[1]=100$，

$c[2]=200$，$c[3]=300$，总财产 E 从 0 开始慢慢增加。

当 E 很小（小于 150）的时候，将 E 平均分给所有继承者，直到各方都拿到 $c[1]/2=50$ 个金币，此时停止给继承者 1 分配财产。

当总财产继续增加（达到 250 个）时，将增加的部分分给剩下的继承者 2 和继承者 3，直到继承者 2 也恰好拿到自己声明值的一半，即 100 个金币，此时停止给她分发财产。

当 E 继续增加（达到 350 个）时，则将增加的部分只分给继承者 3，直到她的损失与继承者 2 相同（即分到了 200 个金币，损失了 100 个金币的时候）。

当 E 继续增加（达到 450 个）时，将增加的部分平分给继承者 2 和继承者 3，直到她俩每人的损失与继承者 1 相同，即每人损失 50 个金币，其中继承者 2 分到 150 个金币，继承者 3 分到 250 个金币。

当总财产 E 多于 450 个金币时，再将继续增加的部分平分给这三个继承者，直到所有人都分得自己所要求的份额为止（共计 600 个金币）。

根据上述分析，我们可以画出按照这一方案分配财产情况的折线图，如图 2.10 所示。

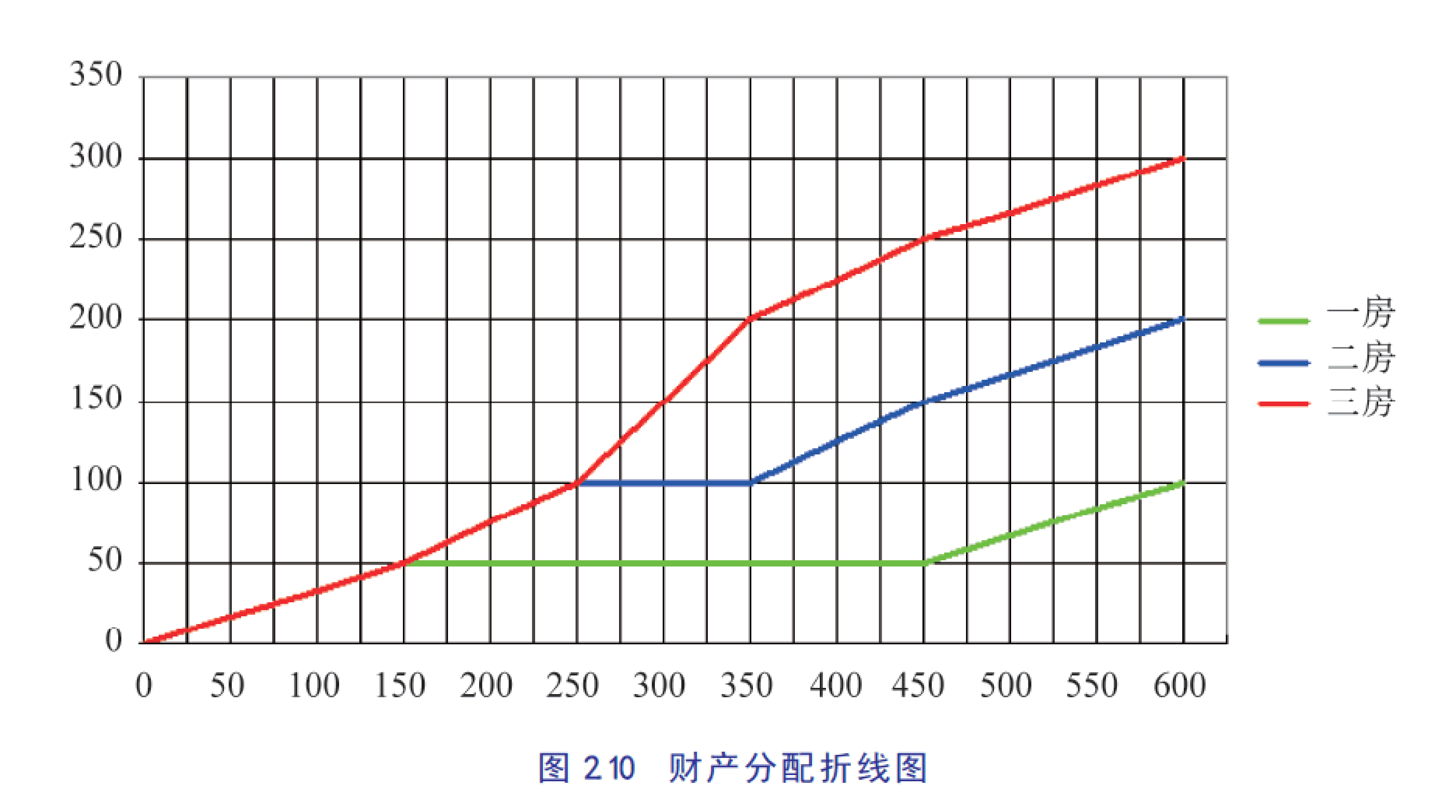


图 2.10　财产分配折线图

可以看到，这一分配方案与《塔木德》中关于“三妾争产”的记载是吻合的。至此，我们终于解决了这一千年难题。

当然，如果这个富翁有更多的妻子，我们可以很简单地把三妾争产问题扩展到 N 妾争产，还是一样的套路，最终的分配方案仍然是倾向于保护弱者的，重点是让最不满意的那位妻子的不满意程度降到最低，换句话说，大家都不太满意，但是我们尽量让最不满意的那位妻子的不满意程度低一些，这样最不满意的妻子看看其他人，觉得大家其实也都差不多，也就接受现实了。在经济学里，这种使得最大的不满最小化的解称为“核仁”解。如果前面提到的“核”是非空集合的话，那么“核仁”一定位于“核”内。

没有绝对的公平

死亡是世界上最公平的事情。

——阿列克谢耶维奇，作家

在分配问题里，公平性是一个最主流的参考标准。为了达到公平性，于是有了 Shapley 值和核。但是，还有些分配中，公平性不再是衡量标准，或者说不可能达到公平性，这时候，应该如何分配呢？

讨价还价的玄机

在社会分配问题中，如果两个人的能力相当，对社会的贡献相等，那么一人一半划分利益，就能得到最大的收益；但如果两个人的能力不对等，对社会的贡献也不相同，此时再公平的划分利益，那么事情的结果就会如同“三个和

尚没水喝”一样，每个人都将失去努力工作的动力。此时，不平均分配才能让每个人有更大的积极性。但同时又要注意不能过于不平均，否则很有可能造成“宁可鸡飞蛋打，也不让你多得”这样两败俱伤、颗粒无收的局面。这说明“做蛋糕”和“分蛋糕”的过程不能完全隔离开，因为分蛋糕的策略还会影响蛋糕最终能做多大。

现在让我们再次来回忆一下前面两节中提到的卖鞋问题和苹果展销的问题。为什么苹果供应商和展销会都能在合作中获得额外利益，而右鞋持有者却在合作中竹篮打水一场空？关键就在于右鞋持有者的谈判能力（或讨价还价能力）为 0，因为有他没他都一样，或者说他完全可以被另外一个人代替。这时，两位右鞋持有者就完全丧失了与左鞋持有者讨价还价的能力，也就是说，他的谈判能力为 0。

当苹果供应商又找了另外一家展销会谈了合作之后，事情也会变得不一样了。试想，供应商与展销会 A 合作，可以卖出 4000 元钱的总价格，而供应商与展销会 B 合作，能创造出 6000 元的总价格。但展销会 B 在不与这家供应商合作时，自己只能赚得 1500 元。那么在供应商与展销会 B 的合作中，展销会 B 的边际贡献为 6000－1500－4000＝500 元。这是因为，即使没有展销会 B，供应商仍然能与展销会 A 合作获得 4000 元的总利润。而供应商的边际效用为 6000－2000－1500＝2500 元。供应商与展销会 B 的边际效用之和为 3000 元，所以供应商要占据 5/6 的比例，而展销会 B 在剩余利润的分配中只能占据 1/6 的比例。剩余利润为 6000－2000－1500＝2500 元，所以供应商分得 2500×5/6＝2083 元的利润，展销会 B 分得 2500×1/6＝417 元的利润。因此，在供应商与展销会 B 的合作中，这 6000 元的总售价，供应商获得 2000＋2083＝4083 元，供应商 B 获得 1500＋417＝1917 元。

注意，虽然供应商 B 的加入成功引诱了供应商，使得展销会 A 退出了与供应商的合作，但正是由于展销会 A 的存在，才使得供应商的谈判能力大幅

提升。如果没有展销会 A 的存在,那么在供应商与展销会 B 的合作过程中,双方的讨价还价能力即会持平,所以会平均分配这 2500 元的利润,供应商只能获得 2000+2500×1/2=3250 元。这多出来的 833 元,就是由于展销会的可替代性太高,致使讨价还价能力低。

因此,为了提升自己在合作中的讨价还价能力,还是应该去尝试与多方进行合作,对比不同的合作关系,才能寻求一种于自身最有利的选择。

分赃要谨慎

下面通过经典的海盗分赃问题,再看一下如何提升自己的讨价还价能力,如图 2.11 所示。

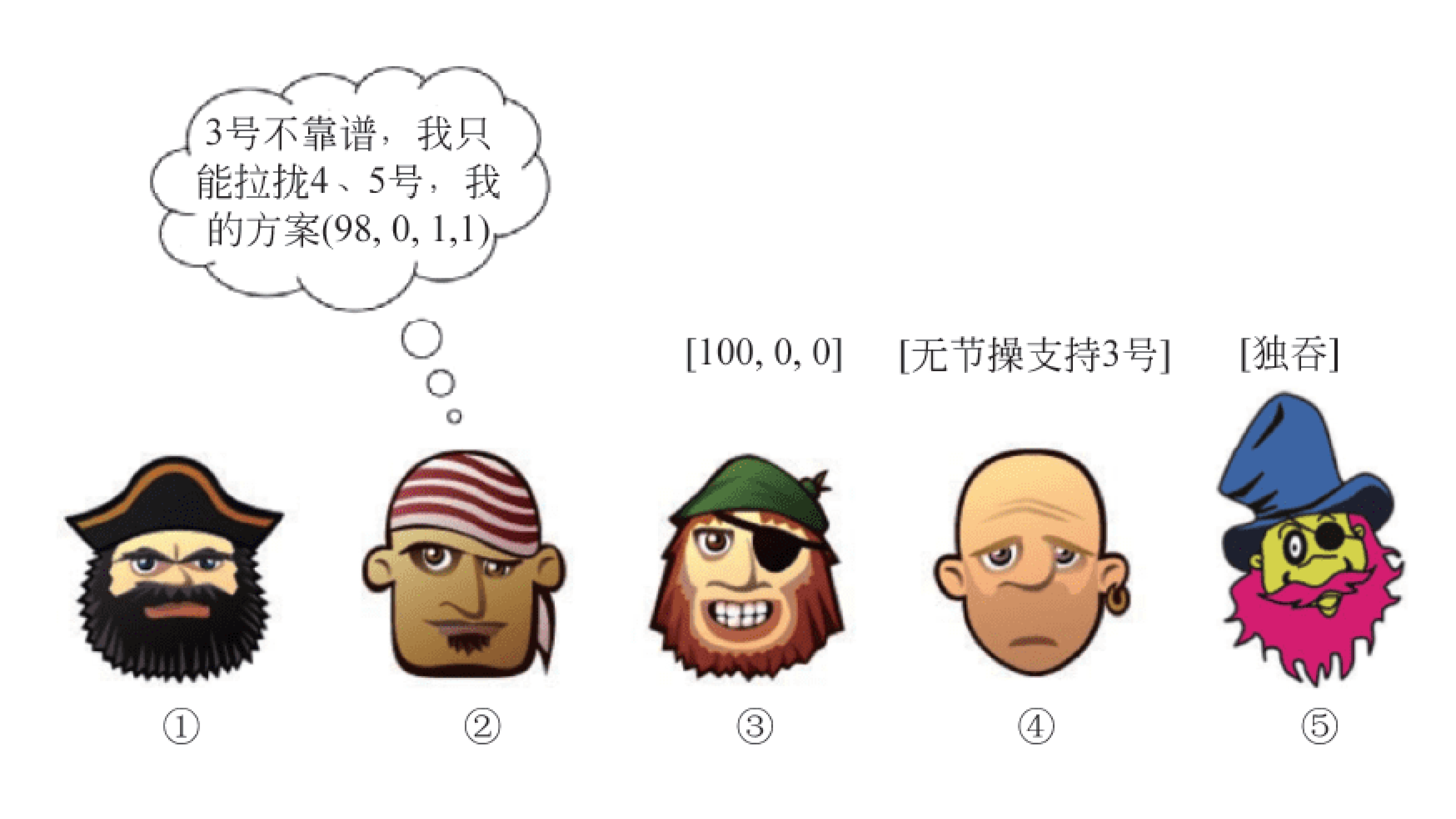


图 2.11　海盗分赃

在索马里有 5 个海盗,某次他们共同抢劫了 100 枚金币,每一枚金币都价值连城。盗亦有道,他们在分赃的过程中决定采取抓阄的办法,具体方法如下:

首先,抽签决定自己的号码:1、2、3、4、5。

之后,由 1 号开始提出分配方案,然后 5 个人举手表决,当且仅当超过半

数的人同意他的分配方案时，才按照他的提案进行分配，若同意方案的人数不足半数，那么他将被扔到大海里喂鲨鱼。

如果1号被喂了鲨鱼，再由2号提出分配方案，剩下4人举手表决。同样，当且仅当超过半数的人同意他的分配方案时，才按照他的提案进行分配，若同意方案的人数不足半数，那么他也将被扔到大海里喂鲨鱼。

以此类推。

若每个海盗都是十分聪明的人，都能很理智地判断得失，从而做出投票选择。那么，海盗们最终的分配结果是什么呢？

这是博弈论的经典案例，如果你是1号海盗，你会平分100枚金币，还是为了保命宁可交出自己的部分？都错了！聪明的1号海盗做出的分配方案是(97，0，1，2，0)或(97，0，1，0，2)。也就是说，1号海盗给自己分了97枚金币，给3号海盗分了1枚金币，给4号或者5号海盗2枚金币。1号海盗为什么敢如此大胆，给自己分得这么多金币呢？为什么他不担心自己的方案得不到其他人的认可，从而会被扔到大海里喂鲨鱼呢？让我们来看一下各位海盗的理性分析。这类问题，一般从最后一个海盗倒推会比较容易得到答案。在博弈论中，这种办法称为逆向推演。

5号海盗：他是最安全的一个，因为没人会把他扔进海里，因此他的想法也十分简单，最好是前面的4个人都死光光，那么他就能独吞这100枚金币了。

4号海盗：他的生存与否完全取决于前面的1～3号是否还活着，因为一旦1～3号都喂了鲨鱼，单独剩下他和5号时，无论他提出怎样的分配方案，5号都一定会投反对票从而把4号扔到大海里喂鲨鱼。即使是4号为了保命提出(0，100)的分配方案把金币都给5号，5号仍然是有可能投反对票的，毕竟留下4号早晚都会是个安全隐患。所以，理性的4号必须要支持3号，才能保住自己的性命。

3号海盗：聪明的3号海盗显然已经看穿了4号的策略，所以他肯定会提出(100，0，0)的分配方案，因为他知道，即使在他的分配方案中4号一无所获，他仍然会无条件地投赞成票，那么再加上自己的一票，3号就能稳稳地把这100枚金币装进自己的口袋了。

2号海盗：他经过推理分析，也料想到了3号海盗的分配方案，所以他决定拉拢4号和5号，而放弃3号的一票，因此他会提出(98，0，1，1)的分配方案。这个方案相对于3号的分配方案，4号和5号至少可以获得1枚金币，理性的4号和5号自然会觉得此方案对他们来说更有利，从而选择支持2号，不希望2号出局而由3号来进行分配。这样，2号可以拿到3票，分得98枚金币。

1号海盗：经过层层推理，1号海盗洞悉到了2号海盗的分配方案，因此他采取的策略是放弃2号这一票而去拉拢3号，因为即使给3号1枚金币，他也会投赞成票的。而对于4号或者5号来说，在2号的分配方案中获得1枚金币，而1号还需要再拉拢一位支持者，因此他需要给4号或者5号2枚金币来诱惑他们中的一人支持自己而非2号。1号再加上自身的一票，97枚金币就能轻松地落入腰包了。因此，1号海盗提出的分配方案是：给自己97枚金币，3号1枚金币，4号或者5号2枚金币。很有意思的结果，对不对？

共享经济平台如何设计最优分配方案

对于国家来说，一个国王的暴政的害处比起不关心公共利益对一个共和国的害处还要小些。一个自由的国家的优点是它的收入分配得比较好，但如果分配得较差的时候，则自由的国家的优点是它根本没有宠臣；但是当事情

不是如此，不是使国王的朋友和双亲发财，而是使参加政府的一切人的朋友和双亲发财的时候，那么一切便都垮台了。

——孟德斯鸠

如今，不管是共享打车，还是共享单车或共享充电宝，共享经济的商业模式不断推陈出新，发展如火如荼，估值也一路飙升。这些共享经济模式凭什么可以这么值钱呢？

共享经济作为新兴经济模式，说到底，最重要的一个创新，就是更好地解决了共享市场中的利润划分问题，因此共享平台的重要性也日益凸显。例如，滴滴公司的存在，不仅在于提供了平台，从而为司机和客户提供了匹配的可能性，更重要的是以一种更为安全和可靠的方式，解决了双方合作的前提——利益分配问题。否则一位私家车车主是无论如何也不会为加班到深夜的你提供顺风车服务的。

任何共享经济平台的设计，都离不开合作机制。网络经济算法在解决这些问题上正在崭露锋芒。

从滴滴想到的一种新型共享经济分配模式的可能性

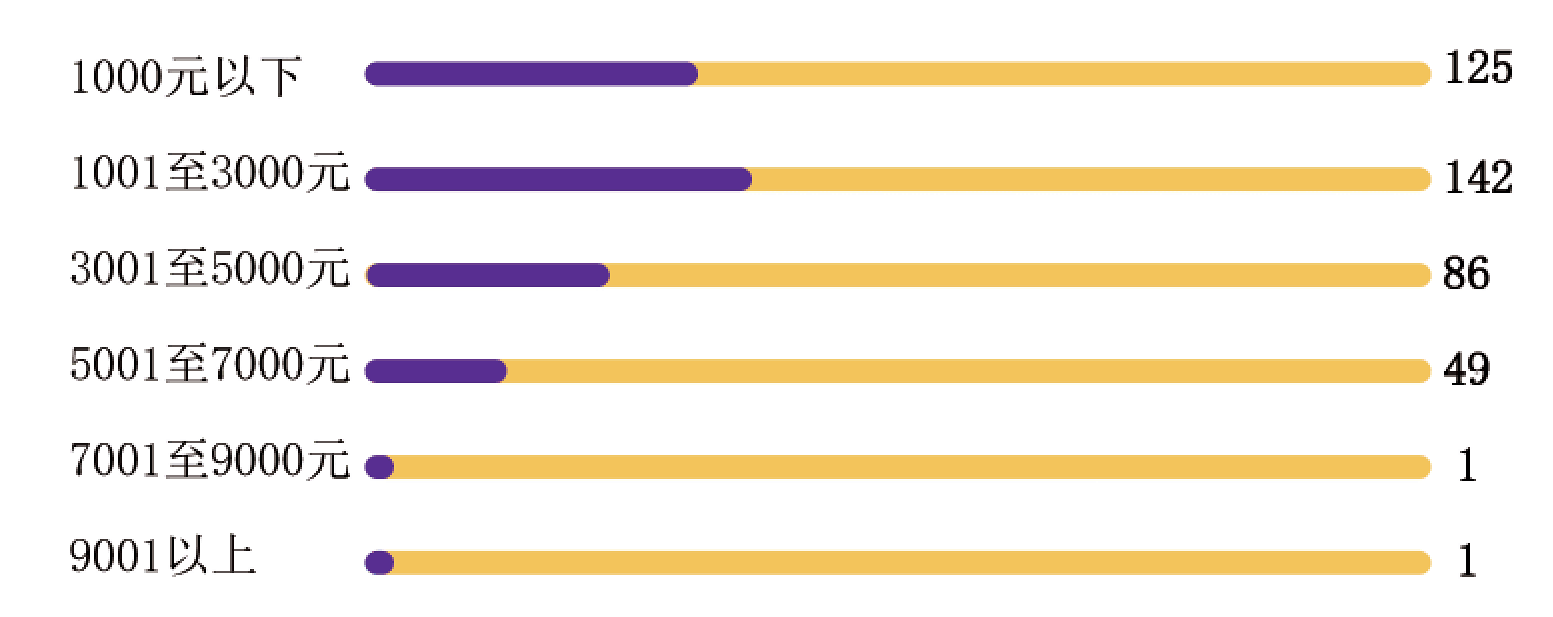
现在专车注册非常方便，不少车主都开起了专车。开专车是一种怎样的体验？其实没有那么多体验生活的感觉，开专车时你就是个司机。真的能赚钱吗？南都记者专车司机修炼秘籍曝光，跑了一整天，赚了 125 块钱。

——南方都市报

2015 年 9 月，南方都市报的一位记者早上 7 点出车，晚上 9 点收车，中间吃午饭和晚饭休息了 3 个小时，当天出车时间共 11 个小时。全天一共接了 20 单，流水费是 314.7 元。根据滴滴专车规定，流水费的两成归滴滴所有，剩下的 8 成归司机所有。除去油钱，剩下的就是利润。由于记者所使用的车

辆较为耗油，百公里油耗为 11 升左右，油钱的成本估计要占到流水费的四成，也就说，记者跑滴滴专车的利润率是 40%。记者一天跑了 11 个小时，利润为 314.7×40 %＝125.88 元。

所以，在你注册成为滴滴司机之前，还是应该首先去了解一下滴滴平台与司机之间的利润划分机制。由于滴滴平台的信息相对难以获得，我们先看一下图 2.12 给出的 Uber 司机的周薪统计信息。



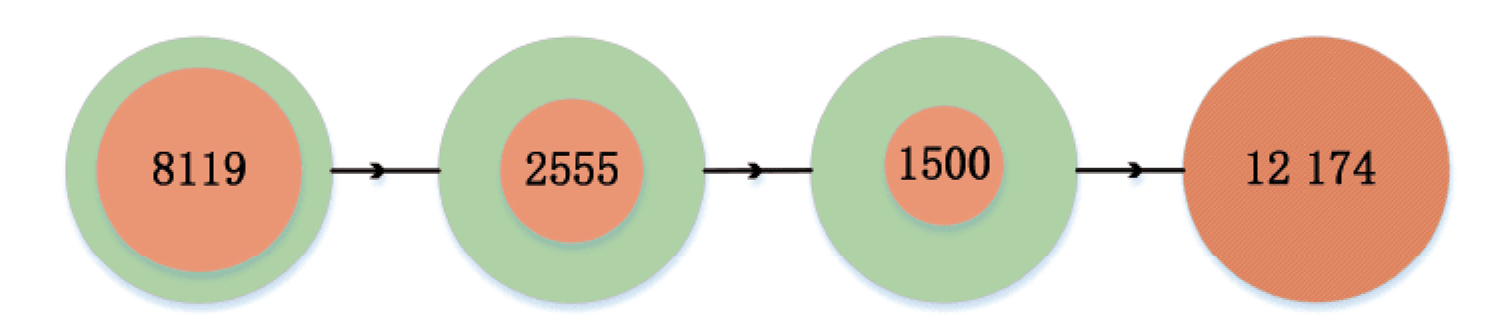


图 2.12　百度滴滴快车吧上给出的 404 位 Uber 司机在 2015 年 7 月的周薪统计

滴滴的利润划分机制

百度滴滴专车吧上的一位司机网友说，滴滴的车费和每单的额外奖励都是按倍数计算的，他以深圳举例，如果当天滴滴是 1.6 倍奖励，那么计算方法如下：如果此单客户支付 15 元，那么你的车费收入就是(15×0.8×0.9832)－0.5＝11.3 元。

为什么要乘以 0.8？因为滴滴公司要先扣 20%。为什么要乘以 0.9832？因为滴滴下属的所谓的中介代理公司要再扣 1.68%（这里注意，这个 1.68% 是滴滴没有注明，也是司机经常算不清楚账目的罪恶黑手）；最后这个 0.5，也就是五毛钱，是滴滴收司机师傅的保险费。话说每单都是如此扣除。

如果暂时忽略对倍数计算和完成单数的奖励，以上就是此单车费计算。

司机对于公司的贡献

2015 年，滴滴公司的估值高达 2280 亿元，与此同时，辛辛苦苦到处奔波一整天的司机只能得到大约 12 元/小时的收入。作为滴滴平台的重要参与方和利益创造者，这样的收入跟滴滴的估值相比，是否匹配？是否公平？要知道，当我们评估这 2280 亿元从何而来的时候，免不了提及这些 12 元/小时的司机们的贡献。

2016 年，由于京沪两地网约车新政变得严苛，快车合规驾驶员数量锐减，而京沪订单大约占到滴滴总量的 15%～20%，驾驶员数量的锐减在一定程度上严重影响了滴滴的业务。2017 年春节前，在线司机数量减少了 25%，一位网名为“滴滴出行高级产品总监”的知乎用户“罗文”指出，由于春节将近，订单数量增长了 30%，致使大量用户吐槽滴滴又贵又难约！

这些问题促使我们思考，是否有一种新的分配形式，可以更好地将参与各方的利益绑定在一起？

如何增加司机的参与感

2016 年，滴滴公司在其公布的《移动出行与司机就业报告》中称，滴滴平台上已经有超过 1500 万司机，覆盖了专车、快车、出租车、顺风车、代驾等多条业务线。司机作为滴滴平台的重要参与者，究竟应该定位为平台的商品还

是平台的价值创造者，恐怕这才是大部分共享经济在商业模式设计中应该考虑的问题，水可载舟亦可覆舟，靠补贴拉来的平台参与各方，一定会在补贴停止时离开。补贴是一个无底洞，虽然在平台建立之初通过补贴可以拉动各方参与，但很显然它不是一种万能药。在持续的竞争中，靠补贴也许暂时可以拖死其他同行，但是一旦补贴停止，就会发现无论客户粘性还是司机粘性，都会迅速消失。

如果排除市场、政策等一系列的因素，我们可以对司机的价值做一个简单的趋势分析：假设 1500 万司机中有 25％流失，会给滴滴公司造成 280 亿元的直接或间接损失，那么这 375 万位司机群体，对于滴滴公司估值的边际效益即为 280 亿元，假设所有司机们的贡献相同，那么每位司机的边际效益为 1866.7 元。

所以，滴滴公司一旦上市，是不是应该给每位司机派送 1866.7 元的股票呢？

当然相比 280 亿，1866.7 元的数字并不高。如果 375 万司机真的能够带来 280 亿元利润的话，1500 万司机能够带来的就不仅仅是 280×4＝1120 亿元这么简单了，而是整个滴滴公司的估值 2280 亿元！如果此时司机们能够团结一致与滴滴公司谈判，那么团结起来的司机数量越大，每个人的所得也就越多。

当然，现实中的问题并不能这么简单地分析，上述过程只是尝试阐述平台与用户之间的合作关系。况且随着参与人数量的增多，团结一致的可能性会急剧下降。但无论如何不应该忘记，共享经济的用户也是平台价值的创造者，因此如何解决合作与分配问题就显得更为重要。我们或许可以借鉴苹果公司的例子，看看共享利润模式如何推动一个公司的价值成长。

分配的多层次幂率效应，赛博新经济体的宿命

赛博新经济体的幂律效应

在 2016 年 9 月苹果公司的新品发布会上，宣布 iPhone 7 的售价分为 5388 元(32G)、6188 元(128G)、6988 元(256G)，iPhone 7 Plus 则为 6388 元(32G)、7188 元(128G)、7988 元(256G)。然而，如果你观察苹果手机的元器件价格，会发现其利润率是非常惊人的。

如图 2.13 所示，苹果公司的全球供应商的利润位于 0.5%～4.7%之间，虽然他们一起参与创造了苹果公司 6000 亿美金的市值，然而与苹果公司 58.5%的利润率相比，可谓是人家吃肉你喝汤。这里面的中国企业也不胜枚举，包括提供镜头红外截止滤光片的环旭电子公司，提供液晶业务的锦富新材公司，提供扬声器模组的歌尔声学公司等。但是，类似这样依附于大平台的合作企业，多数都存在着与此相同的遭遇。

这样的利润分配与幂律分布十分吻合，这并不是苹果公司特有的例子，在任何一个大型合作平台之上，都会存在这种利益分配的巨大差异。只要拥有足够多数量的参与方，那么不管信息对称与否，只要不存在结构性障碍，那么平台上的幂律效应就会尤为凸显。

罗振宇在 2017 的跨年演讲中提到，BAT 三家在 2015 年大约总共有员工 8 万人，这 8 万人一年创造了 2700 亿的收入。也就是行业不到 3%的劳动力，创造了接近行业的一半 GDP。相比之下，余下的这 97%的人创造的人均价值几乎微不足道。

2016 年，中国网络购物用户规模已达 4.6 亿，各种商家店铺也是层出不穷。以商会友网站粗略统计指出，100 多万家的店铺中，皇冠店铺只占到总

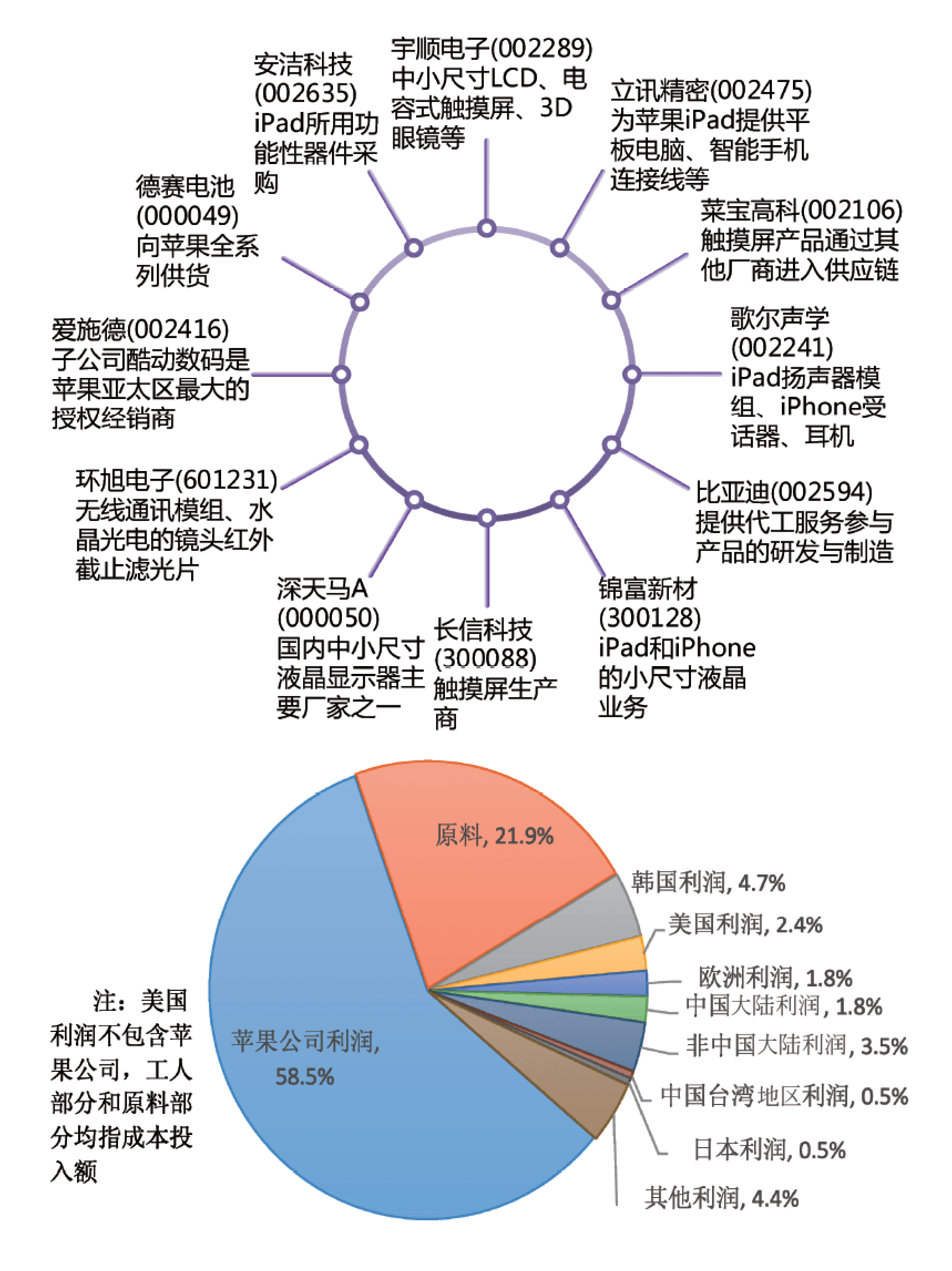


图 2-13　苹果公司产业链及利润划分

数的 0.22%。其中，经营女装/女士精品类的网店数量最多，为 170 819 家，但是皇冠店铺只有 257 家，占其中的 0.15%。由此可见，在诸如淘宝等大型平台上，众多商家之间的客户量和盈利额，基本都遵从着幂律分布。当然，这种幂律效应并不仅仅存在于平台的各个参与方，甚至连平台本身也难逃这一规律。

如图 2.14 所示，2015 年间，在中国的 B2C 市场中，天猫的市场份额位居第一，遥遥领先，京东次之，位列第二，加上大家耳熟能详的苏宁易购、1 号店和亚马逊这几家电商平台，其他所有小平台只占了区区 6.3%的交易总额。

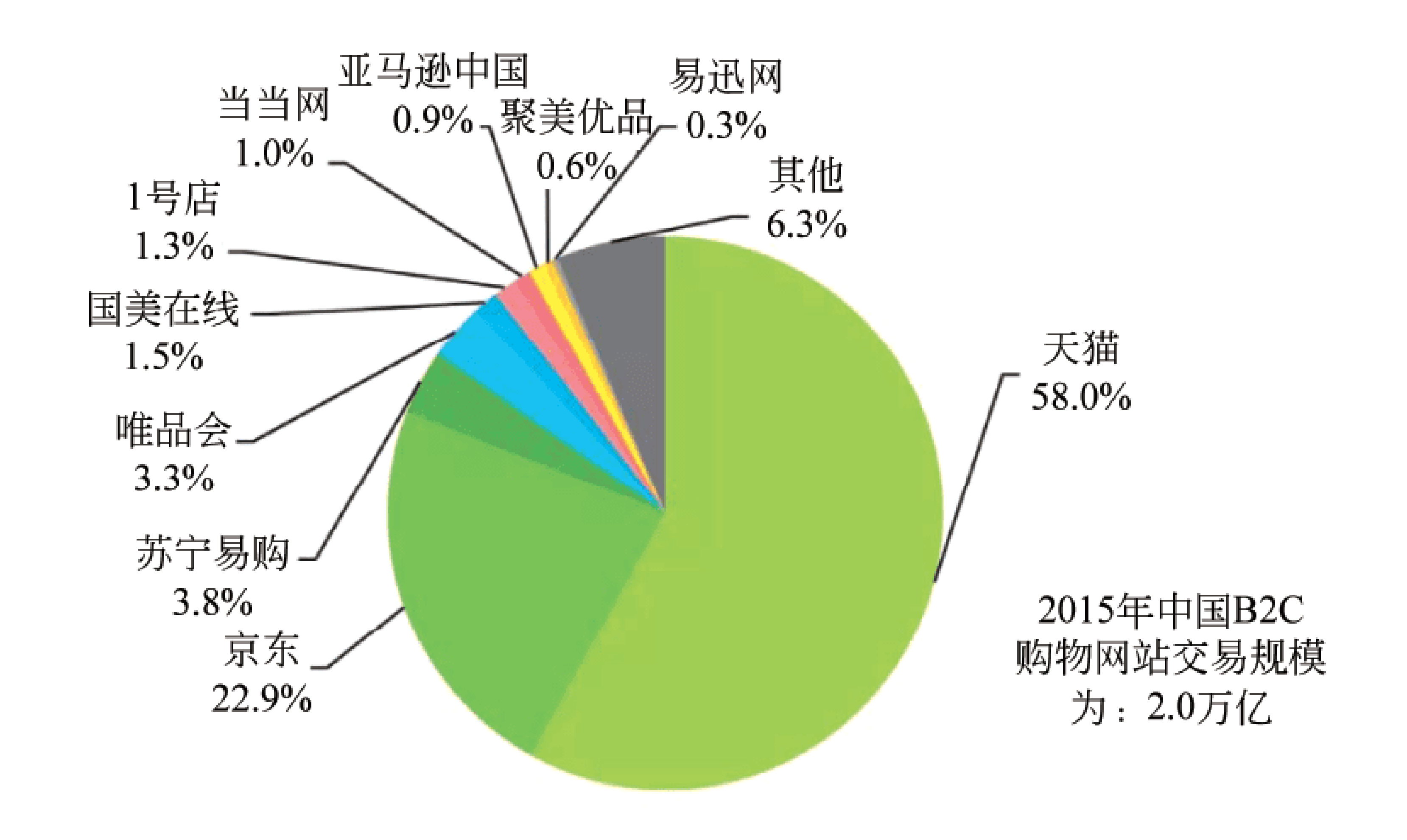


图 2.14　2015 年 B2C 购物网站交易规模市场份额

在 P2P 平台方面，据速途研究院发布的《2014 年 11 月 P2P 市场报告》显示，在全国 1600 家 P2P 平台中，11 月份交易额排名前十的平台累计交易量约为 94 亿元，占行业总成交额的 30%，而红岭创投一家的交易量就为 27.6 亿。

在第三方支付领域，艾瑞咨询的数据显示，早在 2013 年，支付宝以 27.8 亿笔、9000 亿元的支付金额登上了全球最大移动支付公司的宝座，其市场份额也达到了惊人的 78.4%。

幂律效应，简单来说就是指两个通俗的定律，一是二八定律，指出的是不平衡分布，比如 20%最富有的人掌握了全部财富的 80%，二是马太效应，所描述的是穷者越穷富者越富现象。《二十一世纪资本论》用非常可观的篇幅再次向我们阐述了资本获利正在加速的现实，这正是马太效应的直观体现。

幂律其实就是赢家通吃。但这样公平吗？尽管幂律效应早在 100 多年

前就已经由经济学家提出，然而在赛博新经济的环境下，它的效应正在日益凸显，而且这种效应正在由 20∶80 逐渐朝着 0∶100 的方向发展，也就是说，行业引领者将有可能占据接近全部的行业利润，类似于工业经济时代的垄断企业。

在传统经济时代，垄断的存在一直被认为是一种对经济发展和社会稳定不利的商业形态。然而，在赛博经济时代，平台的幂率效应造成的赢家通吃局面，越来越接近于垄断的模式，甚至平台的多边化发展也正渐渐形成一种比垄断更为强大的多边市场的独占经济。

若平台本身既是合作过程的获利者，又是合作规则的制定者，那么它将很难保证平台上各参与方的利益，甚至由于平台的贪婪性，会导致大部分的“长尾”参与方利益受损。我们不能仅仅依靠那种缺乏中立性的平台的自律来替代第三方监管。否则，这种不公平的分配方式将会愈演愈烈，直到破坏新经济的生态平衡。今天，算法为我们建立一个更为公平的赛博市场提供了可能性。我们可以借助算法，更为精确地计算合作博弈的各方利益如何合理分配，这也将为赛博经济新秩序如何建立提供一种新的思路。

第 3 章　匹配算法——双向选择市场里的丘比特

青春小说作家郭敬明写过这样一段文字：

“我相信这个世界上一定有一个你爱的人，他会穿越这个世间汹涌的人群，怀着一颗用力跳动的心，捧着满腔的热和沉甸甸的爱，走向你，抓紧你，他一定会找到你的，你要等。”

这是我们对择偶这件事的文艺化表述。然而，现实总是残酷的，如果你真的这样等下去，很可能你最终一无所获。要知道，虽然“对的人”并不是唯一的，每个单身人士，无论高矮胖瘦，有什么样的职业和爱好，都会有成千上万个“对的人”跟你匹配，令你动心。但是，就算“对的人”成千上万，谁都不能保证必然有一个会穿越汹涌的人群找到你。好消息是，只要你正常地在这个世界上工作和生活，有正常的社交，那么一定存在一个“对的人”。坏消息是，茫茫人海，你如何才能找到他？

如何在网上搜到一个靠谱的女朋友

要想美好地度过一生,就只有两个人结合,因为半个球是无法滚动的,所以每个成年人的重要任务就是找到和自己相配的另一半。

——卡尔·马克思

在加州大学洛杉矶分校的一座教学楼里,35 岁的里斯·麦金利(Chris McKinlay)正试图用 OkCupid(美国最流行的在线约会网站)来找到这个“对的人”。相比现实中的约会,在线交友网站更简单直接,极大提高了人们找到“对的人”的几率。

OkCupid 创办于 2004 年,其联合创始人 Chistian Rubber 毕业于哈佛大学数学系。OkCupid 最吸引交友者的是,他的相亲对象是通过计算方法来自动匹配的。成员通过回答一系列问题进行匹配,比如政治、宗教、家人、爱情观甚至性取向等。在 4000 万通过 Match.com、J-Date、e-Harmony 这些在网络上寻找姻缘的用户中,麦金利当然不会起眼。虽然他已经在网上搜索了 9 个月,可还是毫无结果。他给几十位 OkCupid 网站推荐过来的女士发了邮件,但都石沉大海。

麦金利在波士顿郊区长大,2001 年从明德学院获得了中文本科学位,同年 8 月到纽约世贸大厦做翻译,五周后世贸大楼倒塌那天,他因为提前下班而躲过了一劫。他有点数学上的小天分,在麻省理工学院参加“决战 21 点”时,他一年就赢了 6 万美金。眼下的难题迫使他再次发挥起自己的数学天分。

为了快速收集数据,他一下子建了 12 个 OkCupid 账户,编写了一个

Python 脚本，从这些女士的个人页面上搜集所有可能用到的数据：种族、身高、是否抽烟、星座等。由于只有回答了别人的问题，才能看到别人的信息，麦金利还编写了一个“机器人”（自动化程序）来回答一些简单的问题，以便获取对方的数据。同时，为了避免自己的账号被爬虫系统检测到，他在一位神经学家朋友山姆·托里西（Sam Torrisi）的计算机上安装了间谍软件，根据真人在网站的运动轨迹和数据来设计和改进自己的“机器人”。

这个缜密的女友搜索计划让他在三周内，收集了来自全美国 2 万名女士的 600 万个问题。

这个数据收集办法跟国内的婚恋交友平台如世纪佳缘、百合网等相似，通过大数据建立人物画像，采用聚合算法、神经网络等手段，帮人们寻找最可能配对的人。当你在婚恋平台上更新简历时，网站后台就会运用适当的算法开始帮你物色对象，然后评出分数和匹配度。这些结果都取决于你曾如何回答相关问题，同时通过关联你的社交媒体账号并分析你的行为后所得的数据，比如你在微博、微信等社交媒体发布的内容及更新频率，QQ 登录行为及在线时间，更新图片的频率，作息时间规律，经常出现的位置等，都会成为分析的依据。通过对这些数据做出预测、匹配和推荐，系统就会像亚马逊、淘宝等网站向你推荐你可能喜欢的商品一样，向你推荐你可能会喜欢的人。

这听起来是不是有点不靠谱？通过数据计算的结果，真的能有效匹配吗？要知道，婚恋网站的匹配算法是理性推理的工具，但人是情绪的动物，也常常被情绪奴役。加州大学伯克利分校的一项调查结果显示，美国人只有5%的夫妇是通过网络认识的，个性习惯相同对健康的两性关系帮助并不大。另外，不排除用户会说谎，他们的身高、体重、收入和年龄等都可能有水分。因此，大数据建立起来的模型有时可能偏差很大，需要进行信息修正和过滤。所以，匹配在现实生活中并没有理论所说的那样简单，匹配率最高的人不一定是最理想的交往对象。为了避免这样的现象发生，你还需要再进行甄选的

工作。

所以麦金利随即也开始了他的甄选工作。首先,他根据这些女士的相似性进行分组。这是他在修改 K-Modes 算法时得到的灵感。这个算法最早在 1998 年由贝尔实验室提出,用于分类和分析生病的豌豆谷物。

通过多次参数调整,麦金利根据 2 万名女士的回答将她们分成了 7 个类型。为了验证这些类型是否全面,麦金利又搜集了另外 5000 个刚刚在 OkCupid 上注册的女士样本,这些样本经过处理也大致分布在这 7 个类型里。看起来这个统计方法奏效了。

现在,麦金利需要确定哪个组的女士对自己具有更大的吸引力。有一组女士年龄太小,两组年龄太大,另外一组则是虔诚的基督徒。他还发现有一组女士大多在 20 多岁,多数看起来很独立,有不少从事艺术相关的工作。还有一组女士看起来也很不错,她们年龄稍大,从事编辑和设计等有创造性的工作。麦金利决定在这两组女士中寻找目标,并将其分为 A 组和 B 组。最后,他对两组女士按照匹配率排序,然后开始了他的约会。

虽然有了匹配率排名,可以开始约会,但麦金利仍然面临着一个难题。假设麦金利有 100 名潜在约会对象,他一次只能约会一位女士。那么他是否需要全部约会完才能找到最佳匹配呢?麦金利永远无法知道真正的缘分何时到来,他也不知道下一个约会的女士会是什么样子,停止约会时机实在很难决定。怎么办?

数学家梅里尔·弗勒德(Merrill Flood)在 1949 年提出了“未婚妻问题”。在这个问题中,定义了麦金利需要的决策方案:先依次与部分女士约会,了解自己究竟想要什么,然后再从后面的候选女士中认真考虑谁做自己的未婚妻,具体办法就是与比之前见过的那部分女士都要好的第一个女士发展关系。

从算法上说,就是先拒掉前面 k 个人,不管这些人有多好;然后从第 $k+1$

个人开始，一旦看到比之前所有人都要好的人，就毫不犹豫地选择她。不难看出，k 的取值很讲究，太小了达不到自我认知的效果，太大了又会导致真正可选的余地不多了。这就变成了一个纯数学问题：在约会对象总数已知的情况下，当 k 等于何值时，找到最佳匹配的概率最大。

通过微积分推导，麦金利预计约会者大约有 100 人，他应该先拒绝掉前 37 个人($k=100/e$，$e\approx2.718$ 为自然常数)，静候下一个比这些人都好的人。进一步的计算结果表明，他最终将找到真爱的几率同样是 37%！当然，如果最理想的人在前 37%里，那么错过前 37 个约会对象后，他就再也碰不到更好的人了，因此，他会有 37%的概率“失败退场”，或者被迫选择最后一名约会的女士。同时，我们很容易算出，作为被邀约的女士，最佳的应约时机是在麦金利约会 37 次以后，作为第 38 名约会对象出现。而且，第 100 名也是个好时机，即如果麦金利最理想的人在前 37 名中，此时麦金利“穷途末路”，选中自己的概率至少为 37%。

然而麦金利的潜在约会对象远远超过 100 人。一个多月后，麦金利至少有了 55 次约会。他把每一次约会的记录详细认真地记在笔记本上。约会虽多，但只有三位女士发展到第二次约会；只有一位发展到第三次约会。

大多数约会失败的人都会面临自尊心问题。麦金利也是如此，他甚至开始怀疑自己缜密的计算结果。

夏天快要结束的时候，麦金利收到了一位叫克里斯汀·王(Christine Wang)的 28 岁女孩的留言。根据麦金利的收集的数据显示，这是一位 6 英尺高的蓝眼睛女士，在加州大学洛杉矶分校艺术专业学习，他们的匹配度是 91%。

他在学校的雕塑花园和她见了面，从那里他们走到了学院的寿司店，他立刻就喜欢上了她。他们一起谈论书籍、艺术、音乐，他跟她说了如何通过精心设计的算法找到她的整个过程。

她说："我以为黑客是忧郁的，愤世嫉俗的，但你好像有一点不同。我喜欢这个感觉。"

这是他们总计88次约会中的第一次，接着是第二次、第三次。两个星期后，他们都暂停了他们的OkCupid账户。

这是一个关于匹配的故事。

也许你会说：天，怎么可以用这样冷冰冰的数学公式来计算我未来应该跟谁生活在一起？是的，文学作品总会告诉你，你爱的是一个具体的、活生生的人，而不是一系列指标的集合。然而，麦金利的故事并不是否定人的情感互动，相反，它恰恰证明了在我们存在的这个世界中，婚姻并不是一个一厢情愿的市场，而是一个建立在一系列条件基础上的双向选择。而如何尽快解决在相互选择过程中的有效性，才是事情的关键。就如同诺贝尔经济学奖获得者埃尔文·罗斯(Alvin E. Roth)所说的那样：**我们如何从生活中得到既是我们所选择的，同时也选择我们的事物**。这个寻找人生伴侣的办法，也为我们解决赛博新经济市场中的供需问题提供了一个新的解决方案。

价格在赛博新经济市场的失灵

供求平衡这个概念从19世纪70年代里昂·瓦尔拉斯正式进行分析以来，一直在经济学中发挥着核心作用，稀缺资源的分配问题也是一样。从这个意义上来说，肾脏移植匹配属于经济学的研究范围之内。

——坂井丰贵，《合适》

20年前，巴黎将地铁车厢分为一等车厢和二等车厢，然而，让人们不解的是，两种车厢的座位数目和质量完全相同，唯一的区别是一等车厢的价格

是二等车厢的两倍。这种看似不合理的定价机制，却获得了更高的客户满意度。这种定价策略利用不同人群对价格敏感度的不同，通过设计不同的产品来匹配不同用户的需求，那些对价格敏感度低、同时又注重舒适度的乘客可以选择相对宽松的一等车厢，价格敏感度高对舒适度不敏感的乘客可以选择相对拥挤的二等车厢。

这就是著名的“巴黎地铁定价”方案，其背后的逻辑正是人们对价格的不同敏感程度。大部分的普通人会选择便宜的二等车厢，因此，二等车厢通常都是拥挤的。而那些更注重舒适度的人士，更倾向于花更多的钱进入相对宽松的一等车厢。

“巴黎地铁定价”是一个典型的利用价格控制资源分配的例子，类似的例子还有最近出现的网约专车的计价机制。如果我们不考虑叫车平台对用户的补贴，那么通过专车平台叫到的汽车一般会比出租车贵，以满足那些对品质和服务有要求的客户。不仅如此，在用车高峰时段需求激增时，这些公司通常还会采取动态提高价格的策略①。

在一般的商品市场中，价格作为“看不见的手”，决定了谁能得到什么。你可以在成千上万的淘宝店铺里为自己挑选产品，只要你买得起，你就一定可以得到你想要的东西。在这里，价格负责了一切，你只需要让自己变成那个“买得起”的人。

然而，在某些市场里，你会发现钱并不是万能的，价格并不能决定一个人得到他想要得到的东西。这类市场通常有两个特点：首先，价格对需求的影响是有限的。同样是交通运输，将“巴黎地铁定价”机制搬到人口数量占据全球将近五分之一的中国，动态调整价格带来的惩罚和抑制作用在“刚需”面前很可能失效。2001年至2007年实行的春运火车票价格浮动非但没能缓解

① 本书第4章会详细介绍动态定价的机制和算法。

春节出行压力，反而带来了更大的社会矛盾。其次，在某些场景下，价格难以发挥其本来的作用。例如，名校的就读名额一般情况下是不能出售的。在职场中，公司也不会通过不断地降低工资来寻找求职者，恰恰相反，公司会用丰厚的薪资或福利来吸引高水平的人才，并选择最优秀、最忠诚的员工。

因此，学生与学校、雇主与求职者之间都存在双向选择的问题，从而形成一个与传统商品市场不同的新型市场，在这个市场中，传统的市场机制很难确保供需双方之间实现更好的匹配。高校录取和劳动力市场匹配实质上是对不可分的离散资源（学生和高校、雇主和求职者）进行配对，就像男女婚姻匹配一样，是一个**相互取悦对方并做出选择的双向匹配市场**。其实，不只是这些，合租、器官移植、移动通信牌照和频段分配等都面临相类似的匹配问题，其主要特点在于难以通过价格机制实现市场调节功能。

因此，在传统的商品市场之外，还存在一种需要买家与卖家之间互动匹配来达成交易的新型市场，我们称之为双向匹配市场。在双向匹配市场中，商品（实际上，这里说商品不太准确，我们暂且先用这个词）可以按人们的喜好分配，但是没有明确的买入、卖出或价格设定。例如，可供移植的肾脏不够分配或者最好的学校名额不够时，这些稀缺资源必须要通过遵循某种规则进行配置。

综上所述，我们发现双向匹配市场有两个重要的特征：第一，参与匹配的双方分别属于两个互不相交的集合，且位置不能互换。例如，高考志愿填报中，学生和高校已确定且无法互换。第二，只有经过双方一致同意后，才能形成匹配，即匹配是双向的。在男女婚姻中，落花有意流水无情往往无法修成正果。同样，再好的大学也不可能招到所有优秀的学生，还需要这些学生也正好选择了这些学校。劳动力市场亦是如此，这里，我们暂不考虑求职者自主创业进一步成为雇主的可能性，即求职者和雇主这两个集合不存在交集。

在传统的商品市场中，价格是亚当 · 斯密所说的“看不见的手”。而**在匹配市场中，面对价格的失灵，让买方和卖方可以找到对方的匹配算法就发挥了其重要的作用。针对不同的匹配需求，市场设计有不同的侧重点，有的追求稳定匹配，即双方都不可能在这个市场中找到更好的选择，有的追求最大匹配，即在市场中能够达到匹配的参与者数量最大，相应地，我们也将使用不同的匹配算法，即稳定匹配算法和最大匹配算法。下面我们就以这两种算法的典型应用来介绍它们如何在市场中发挥作用。**

Gale-Shapley 匹配算法：一种更为稳定的匹配设计

男女关系方面，社会就那么多资源，基本上会趋向公平合理的分配，只有很少一部分人会投机成功，最终，多数人还是会找到合适的，如果自己觉得不合适，那是自我认知有问题——大家都想找比自身条件更高一级的，不过最后总会落到同等水平的人手里，自由在哪里？我认为是向低一档的人群里找。

——石康，作家

自己喜欢的人恰好也喜欢自己，然后共结连理，这大概是最理想的婚姻了。不过人生总是如此艰难，落花有意流水无情才是经常上演的戏码。然而，这并不是最悲催的，如果婚后才发觉遇到了真爱，恐怕才是悲剧的开始。一个稳定的婚姻应该是建立在双方都遵守婚姻规则的基础上，彼此都能够主动或被动地避免出轨行为。通俗地说，在婚姻内，如果**“我爱的人也爱我”**或者**“我爱的人不爱我，爱我的人我不爱”**，都可以形成稳定的婚姻。

也许这会让一些人听起来不太舒服，毕竟在他们的价值观里，爱情或婚姻中的精神与肉体双重统一是他们一直以来的信仰，但是我们如果以结果为判断依据时，就会发现事实的确如此。如图 3.1 所示，比如说李先生和孙女士婚后，李先生虽然又喜欢肖女士，但肖女士并不认为李先生与自己合拍，她喜欢的还是王先生，那么李先生和孙女士的婚姻关系就会处在相对稳定的状态，这时我们才能说这个匹配是有效的。也许社会学或者心理学家会给你许多有关匹配的思考和哲学，然而罗伊德·沙普利（Lloyd Shapley）和戴维·盖尔（David Gale）却用算法来告诉你一个可以具体操作的办法，那就是他们提出的 Gale-Shapley 匹配算法。



图 3.1　稳定的婚姻匹配

Gale-Shapley 匹配算法

匹配算法的灵感最初来源于一封如何选择室友的信件。罗伊德·沙普利收到普林斯顿的同班同学之一、加州大学伯克利分校教授戴维·盖尔（David Gale）的一封信，他提出了选择室友的问题：如果有两组人，每个人都有不同的偏好，是否有方法产生一个将两组人配对的稳定解？盖尔起初认为没有方法可以产生一个稳定解，然而当沙普利试图证明这个结论的时候，却意外地发现这个问题存在稳定解！后来，他与盖尔将这个算法写成了一篇名为“大学招生和婚姻稳定”的论文，发表在 1962 年的《美国数学月刊》上。这就是著名的 Gale-Shapley 匹配算法。然而，他们的论文最初并未被评审人接

受，它被诟病的地方是一篇数学论文竟然没有包含任何数学方程。对此他们在论文中是这样解释的：

“推导过程没有晦涩的辞藻或技术术语，我们只通过通俗的语言代替数学符号来描述算法，因此并不需要微积分的知识作为前提。实际上，一个人只要会计数就可以。然而任何一个数学家很快就能意识到这的确是数学推导。”

这个天才的算法后来被哈佛大学商学院教授埃尔文·罗斯（Alvin E. Roth）发扬光大，他在 Gale-Shapley 机制的基础上，开发了一种实用算法，并将其应用在了很多领域，比如美国的住院医师、职业劳动市场，以及纽约和波士顿的公立学校招生等。这极大地推动了 Gale-Shapley 匹配算法的应用和发展。

2012 年，瑞典皇家科学院将该年度诺贝尔经济学奖授予了罗斯与沙普利，以此嘉奖两位学者在稳定分配理论及其市场设计实践方面的成就。瑞典皇家科学院陈述两位教授获奖理由为：“稳定分配——由理论走向实际”。正如前面所说，沙普利是理论的先驱，而罗斯则是一个实践者。遗憾的是盖尔教授于 2008 年溘然长逝，不然他一定会与沙普利分享诺贝尔奖的崇高荣誉。

经典的 Gale-Shapley 匹配算法（后面简称为 Gale-Shapley 算法）是一个迭代算法：第一步，每个单身男士在所有没有拒绝他的女士中选择最喜欢的求婚；第二步，每个女士在所有向自己示爱的男士中选择最喜欢的作为男友。如此迭代，直到每一位男士要么牵手一位女士，要么被所有女士拒绝。注意，这里面有一个假设的前提条件是，被拒绝一方不会继续执着地追求对方，虽然实际情况往往不会如此（抱歉在使用数学方法时，我们总是需要很多理性！）。

假设有 4 男 4 女，男女双方都有一个偏好列表。例如，1 号男士对女士的偏好依次为 3 号女士、4 号女士、2 号女士和 1 号女士。同样，每位女士也

对男士们有一个偏好列表。每位男士都向最心仪的女士发起攻势，向她表白。女士将面对三种情况：有不止一个人跟她表白；只有一个人跟她表白；没有人跟她表白。这三种情况对应的选择分别是：选择最优选的那一位，答应与他暂时配对；接受表白，暂时与其配对；继续等待。当此轮选择结束后，男女均有一部分已完成配对，剩下的继续保持单身。如图 3.2 所示，在第一轮中，1 号男士和 2 号男士同时向 3 号女士表白，但是假设 3 号女士更喜欢 1 号男士(女方的偏好在图中未列出来)，因此，2 号男士会被拒绝，在这一轮结束后，2 号男士仍然是单身。

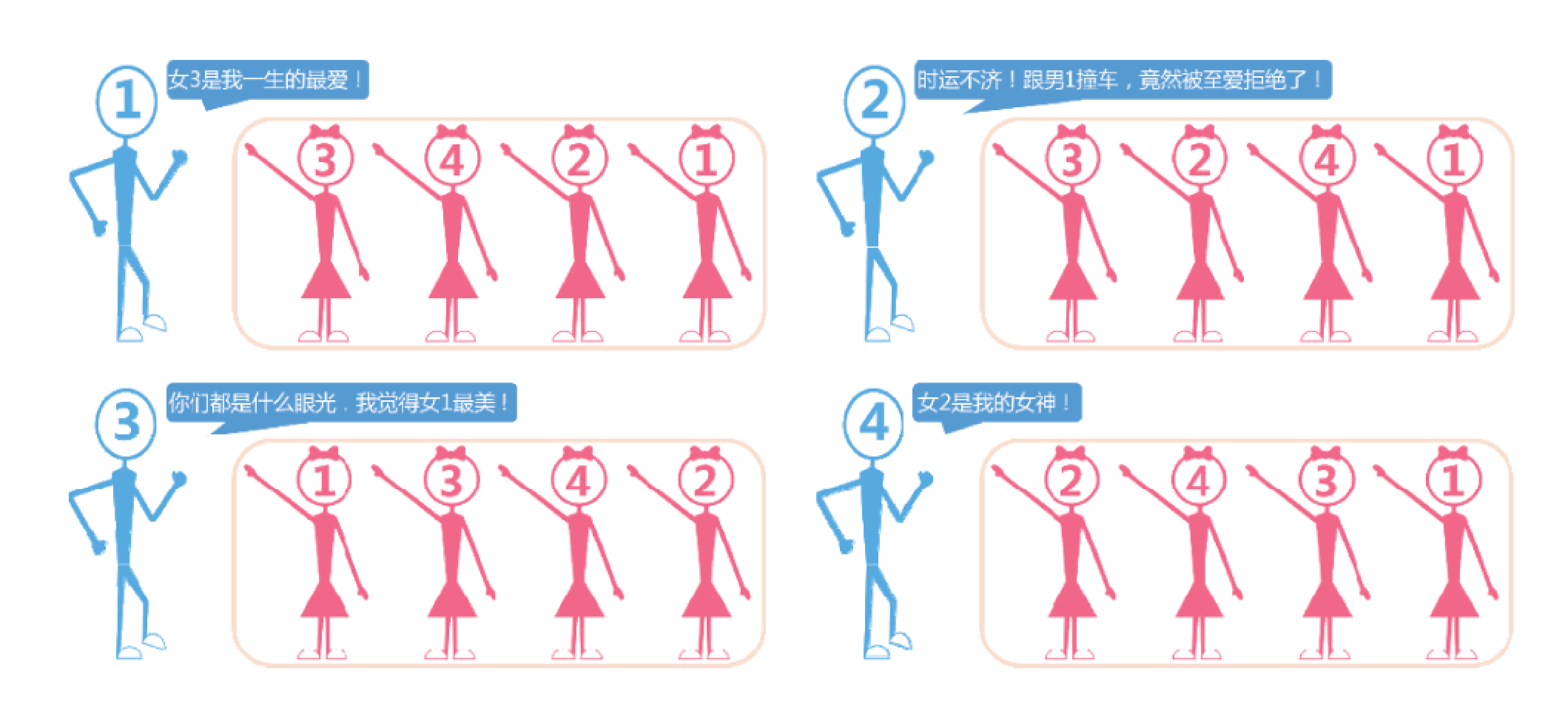


图 3.2　男士偏好列表示例（第一轮表白）

只要匹配没有最终完成，这个过程就会继续，在新一轮中，每位未配对男士都像前一轮一样，女士们则需要从表白者中选择最中意的一位，拒绝其他追求者。因此，如图 3.3 所示，在第二轮迭代中，2 号女士为 2 号男士的第二选择，且在此前没有拒绝过 2 号男士，故 2 号男士直接向 2 号女士表白。而在 2 号女士的排序中 2 号男士优于 4 号男士，因此 2 号女士必须拒绝 4 号男士，与 2 号男士配对。4 号男士重新成为单身。

继续上述过程，如图 3.4 所示，4 号男士选择向 4 号女士表白，而由于 4 号女士一直处于等待状态，她会接受配对，于是配对稳定下来。男士拥有主

动选择权，但是男士经过一轮一轮筛选，追求对象却越来越差，女士虽被动，却可以得到相对更好的选择。配对完成后，男女双方均得到规则下相对最好的选择，成为稳定配对，如图 3.5 所示。

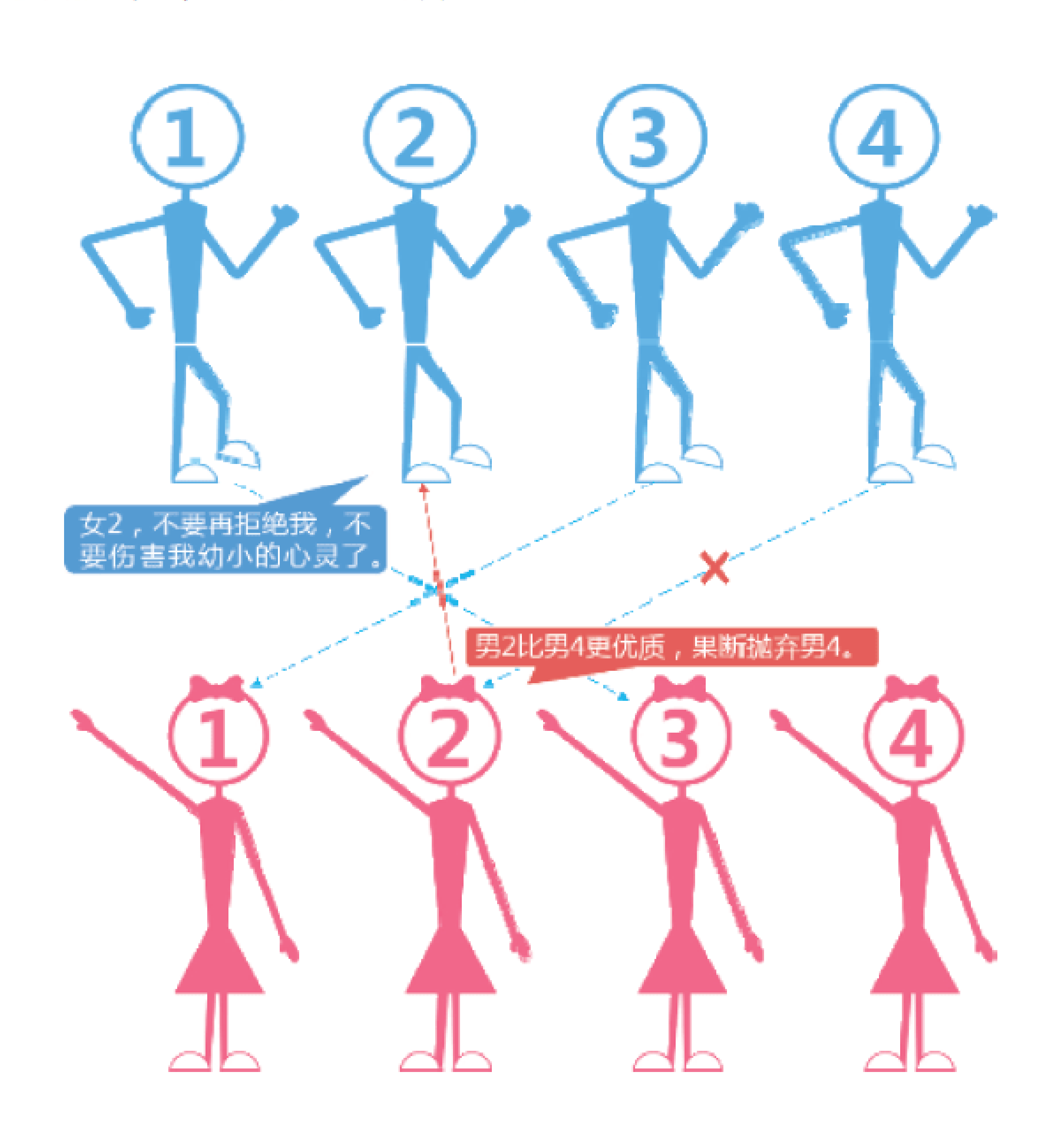


图 33　第二轮迭代

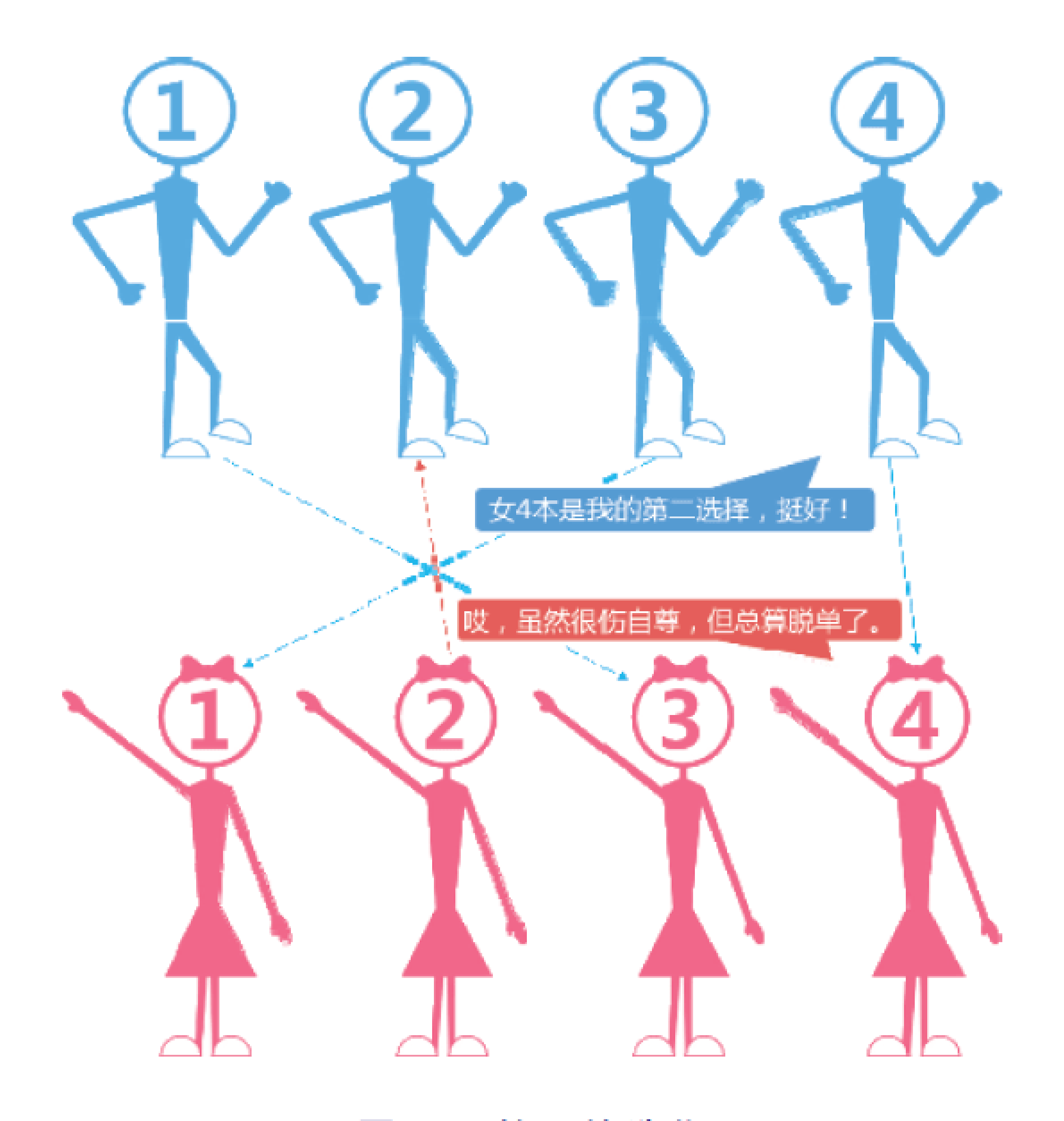


图 34　第三轮迭代

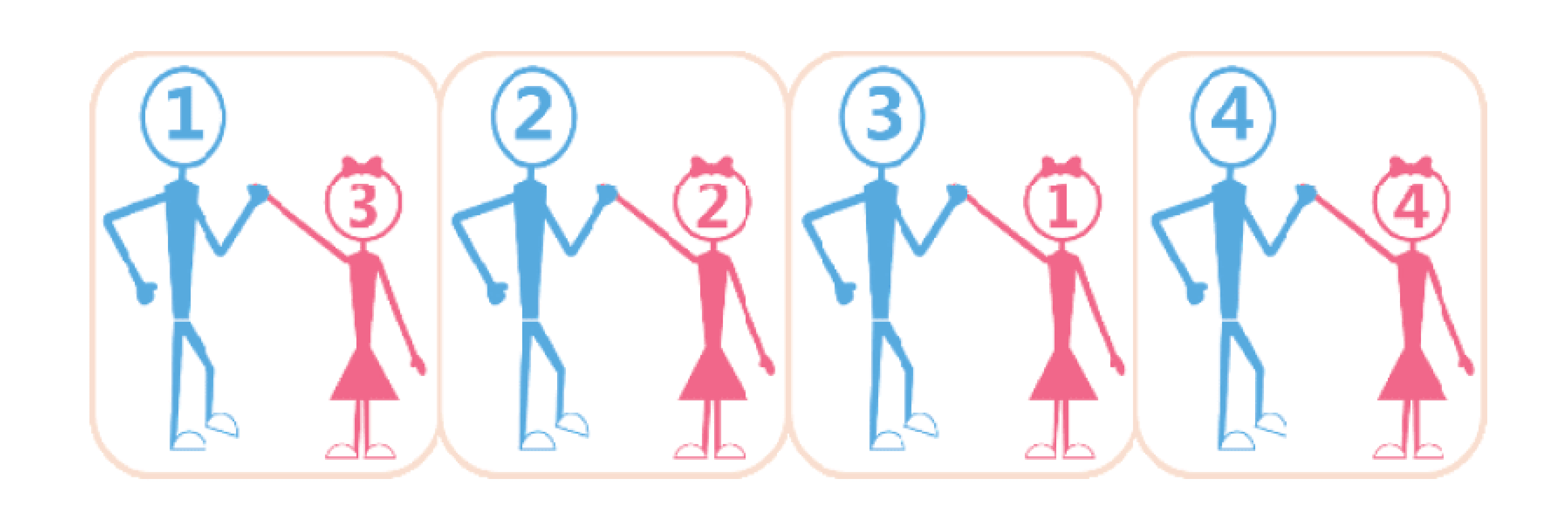


图 35　最终配对结果

这个世界存在稳定匹配吗

也许有人会质疑 Gale-Shapley 算法的有效性，这个如此简单的算法是如何解决稳定性问题的？为了解答这个疑问，首先我们需要达成共识，即互相选择的双方在数量相等的情况下，最终将全部完成互相配对，不会有人落下。在这个前提条件下，Gale-Shapley 算法得到的匹配结果如何确保是稳定匹配呢？究竟还有没有不稳定因素呢？我们用下面的逻辑进行一个简单推理：

假设匹配结果中存在不稳定因素。有一男一女，分别是 M 和 W，他们各自都已经有了伴侣，但是 M 喜欢 W 胜过他现在的伴侣，同样，W 也喜欢 M 胜过她现在的伴侣。但是根据算法规则，M 肯定是向 W 求过婚的，如果 W 更喜欢 M，W 应该选择 M 而不是当前的伴侣，所以这个假设不会成立，也就是说，这种不稳定的因素并不会发生。

可见，Gale-Shapley 算法迭代的结果最终一定会达到一个稳定均衡点。这个均衡点就是一个稳定匹配，使得任何一对配对的男女没有动力背叛对方，这相当于非合作博弈中的纳什均衡。

虽然大部分被歌颂的爱情都在标榜执着地追求命中注定的唯一，但在匹配理论中，**稳定的匹配，即“对的人”并不是唯一的**，而且稳定的匹配也不代表每

个人都能得到自己最梦寐以求的人，甚至在一个稳定匹配中，没有人能够与自己最喜欢的人在一起。赛博新经济市场的匹配也是如此。这听起来或许令人沮丧，但这并不代表你应该放弃执着地追求。相反，你会从中发现主动与被动地去参与匹配会得到不同的结果。这就是匹配算法常提到的两个策略，即“男士优先”还是“女士优先”，这两种策略导致的结果可能有天壤之别。也就是说，Gale-Shapley 算法并不是对称的。前者男士占据主动：如果还存在其他的稳定匹配，那么里面任何一个男士的伴侣排名都不会比“男士优先”得到的结果更好，我们说此种情况每个男士获得的是“最好”的伴侣。此时女士只能被动地一步步接近她最爱的目标，也许最终往往碰不到，结果是：如果还存在其他的稳定匹配，那么里面任何一个女士的伴侣排名都不会比“男士优先”得到的结果更差，我们说此种情况每个女士获得的是“最差”的伴侣。第二种情况恰好相反。在匹配问题中，这种现象称为“利益的对立”（Opposition of Interests）。盖尔在 1985 年进一步证明了：“所有的双边匹配问题中，双方必定存在利益的冲突”。这种利益冲突会导致在婚姻中双方不会拥有同等的满意度。所以，匹配算法也会用数学的方法告诉你，**主动追求幸福的人，会更加接近幸福**；主动参与选择的人，更可能得到对自己最有利的匹配结果。

匈牙利算法：一支最大匹配的丘比特之箭

在黄石谷谈得来，不一定在费城也谈得投机，在大城市中，有着太多转移我们心思的因素，我老觉得一男一女流落在荒岛上，立刻可以结合，因没有选择的缘故。在城里，有选择的时候，男女间感情发展往往是比较缓慢的。

——亦舒，作家

当然，在匹配市场里，并不是所有的匹配都是以解决稳定性为首要目标，有时候，相比稳定性，市场双方更追求最大化地实现配对，因此在经典的Gale-Shapley算法基础上，出现了针对这一匹配目标的匹配算法。我们用"丘比特的烦恼"为例来介绍，这是2010年国际信息学奥林匹克竞赛中国队选拔赛(China Team Selection Competition，CTSC)中的一道题目。

丘比特的烦恼

随着社会的不断发展，人与人之间的感情越来越功利化。最近，爱神丘比特发现，爱情也已不再是完全纯洁的了。这使得丘比特很是苦恼(见图3.6)，他越来越难找到合适的男女，并向他们射去丘比特之箭。于是丘比特千里迢迢远赴中国，找到了掌管东方人爱情的神——月下老人，向他求教。



图3.6　丘比特的烦恼

月下老人告诉丘比特，纯洁的爱情并不是不存在，而是他没有找到。在东方，人们讲究的是缘分。月下老人只要做好一男一女两个泥人，在他们之间连上一条红线，那么它们所代表的人就会相爱——无论他们身处何地。而丘比特的爱情之箭只能射中两个距离相当近的人，选择的范围自然就小了很多，不能找到真正的有缘人。

丘比特听了月下老人的解释，茅塞顿开，回去之后用了人间的最新科技改造了自己的弓箭，使得丘比特之箭的射程大大增加。这样，射中有缘人的机会也增加了不少。

情人节(Valentine's day)的午夜零时，丘比特开始了自己的工作。他选

择了一组数目相等的男女，感觉到他们互相之间的缘分大小，并依此射出了神箭，使他们产生爱意。他希望能选择最好的方法，使他所选择的每一个人被射中一次，且每一对被射中的人之间的缘分之和最大。

当然，无论丘比特怎么改造自己的弓箭，总还是存在缺陷的。首先，弓箭的射程尽管增大了，但毕竟还是有限的，不能像月下老人那样，做到“千里姻缘一线牵”。其次，无论怎么改造，箭的轨迹终归只能是一条直线，也就是说，如果两个人之间的连线段上有别人，那么就不能向他们射出丘比特之箭，否则，按月下老人的话，就是“乱点鸳鸯谱”了。

“作为一个凡人，你的任务是运用先进的计算机为丘比特找到最佳的方案”。

这里，CTSC 让大家运用的计算机方法其实就是一个有关匹配的算法。这个问题与经典 Gale-Shapley 稳定婚姻问题有所区别：

首先，弓箭的射程有限。所以某一男士只能在有限的范围内寻找自己的伴侣，不可能像在稳定婚姻问题里那样，可以在所有女士范围内选择。

其次是匹配的目标不同。丘比特的目标是找到缘分值之和最大的匹配，而稳定婚姻追求的是匹配的稳定性。可以说，**稳定婚姻匹配是发散的，具有博弈的特性，而“丘比特之箭”是集中式的控制**。

稳定婚姻可以采用 Gale-Shapley 算法来解。而丘比特的烦恼则需要新的匹配算法来解决。首先，我们假设男女之间缘分值是没有区别的，那么该问题就变成了最简单的最大匹配问题。也就是一个二分图最大权匹配问题。二分图指的是这样一种图，所有的边都在图的两个节点集合之间，而两个节点集合内部没有边。那么，我们需要做的就是尽可能地为最多的节点实现配对，寻找二分图的最大匹配一般采用的是匈牙利算法。

匈牙利算法

匈牙利算法由匈牙利数学家 Edmonds 于 1965 年提出，因而得名。如图 3.7 所示，假设有 4 个剩男（左），4 个剩女（右），每个人都可能对多名异性有好感。如果一对男女互有好感，那么就可以把这一对撮合在一起。在图 3.7 所示的好感关系图中，每一条黑色实线都表示互有好感。匈牙利算法就是尽可能地撮合更多的情侣，也就是实现二分图的最大匹配。

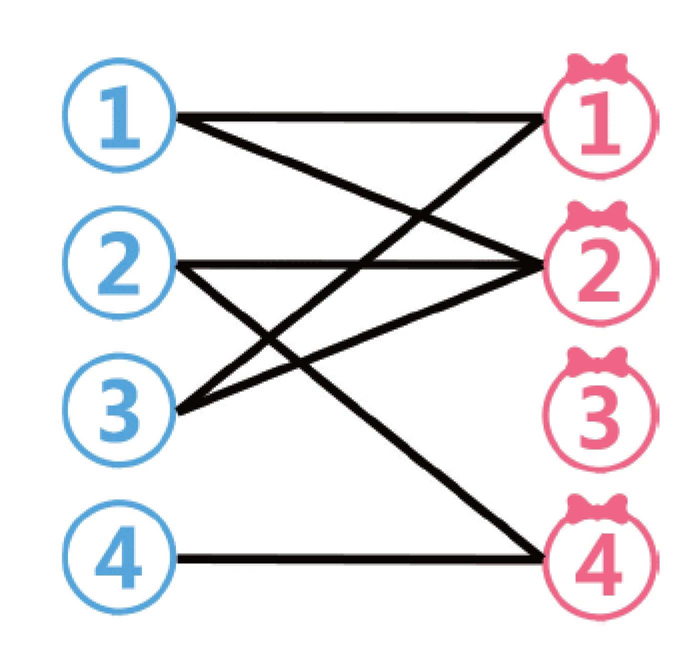


图 3.7　好感关系图

第一步：先试着给 1 号男士找女士，发现第一个与他相连的 1 号女士还单身，此时，连上一条红线，如图 3.8 所示。

第二步：接着给 2 号男士找女士，发现第一个与他相连的 2 号女士也单身，同样，连上一条红线，如图 3.9 所示。

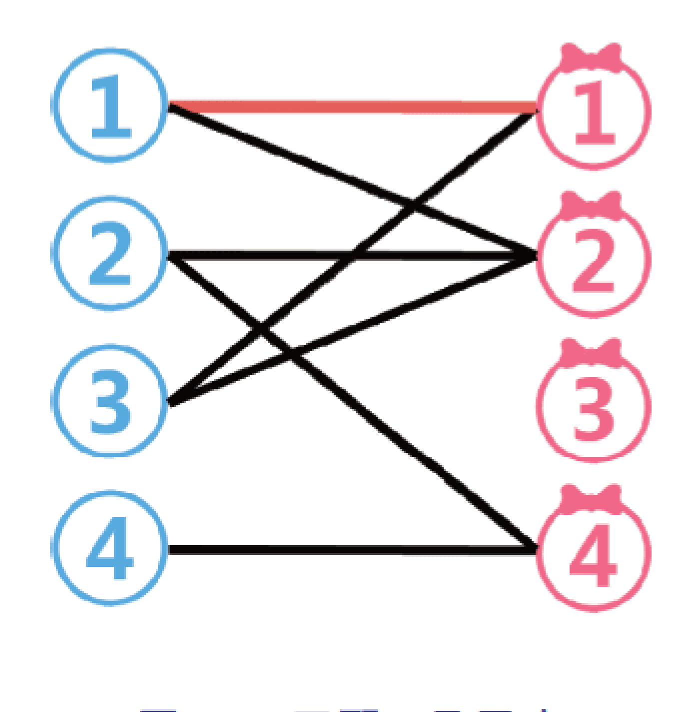


图 3.8　匹配 1 号男士

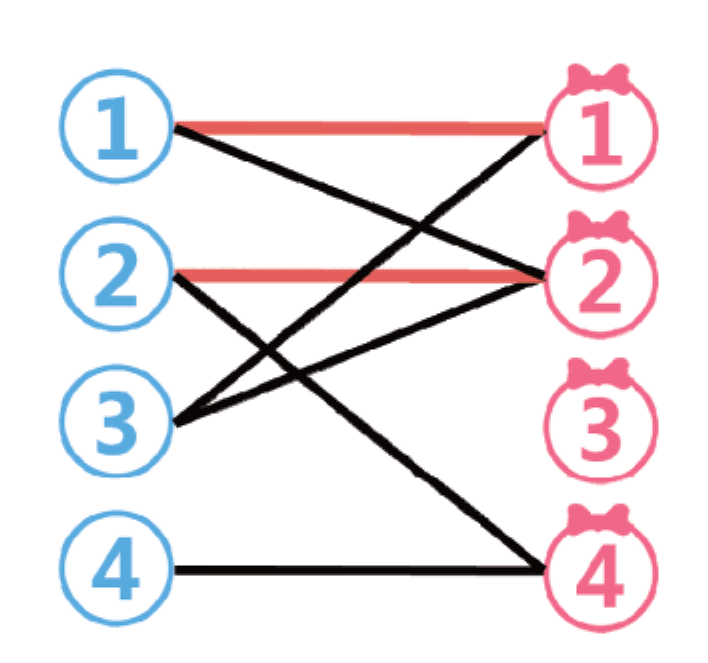


图 3.9　匹配 2 号男士

第三步：接下来是 3 号男士，很遗憾 1 号女士已经有主了，怎么办呢？我们试着给之前 1 号女士匹配的男士（也就是 1 号男士）另外分配一个女士（用虚线表示，说明这条边被临时拆掉），如图 3.10 所示。

与 1 号男生相连的第二个女士是 2 号女士，但是不巧的是 2 号女士也有对象了，怎么办呢？我们再试着给 2 号女士的原配（2 号男士）重新找一位女士（注意这个步骤与上面是一样的，也是一个递归的过程），如图 3.11 所示。

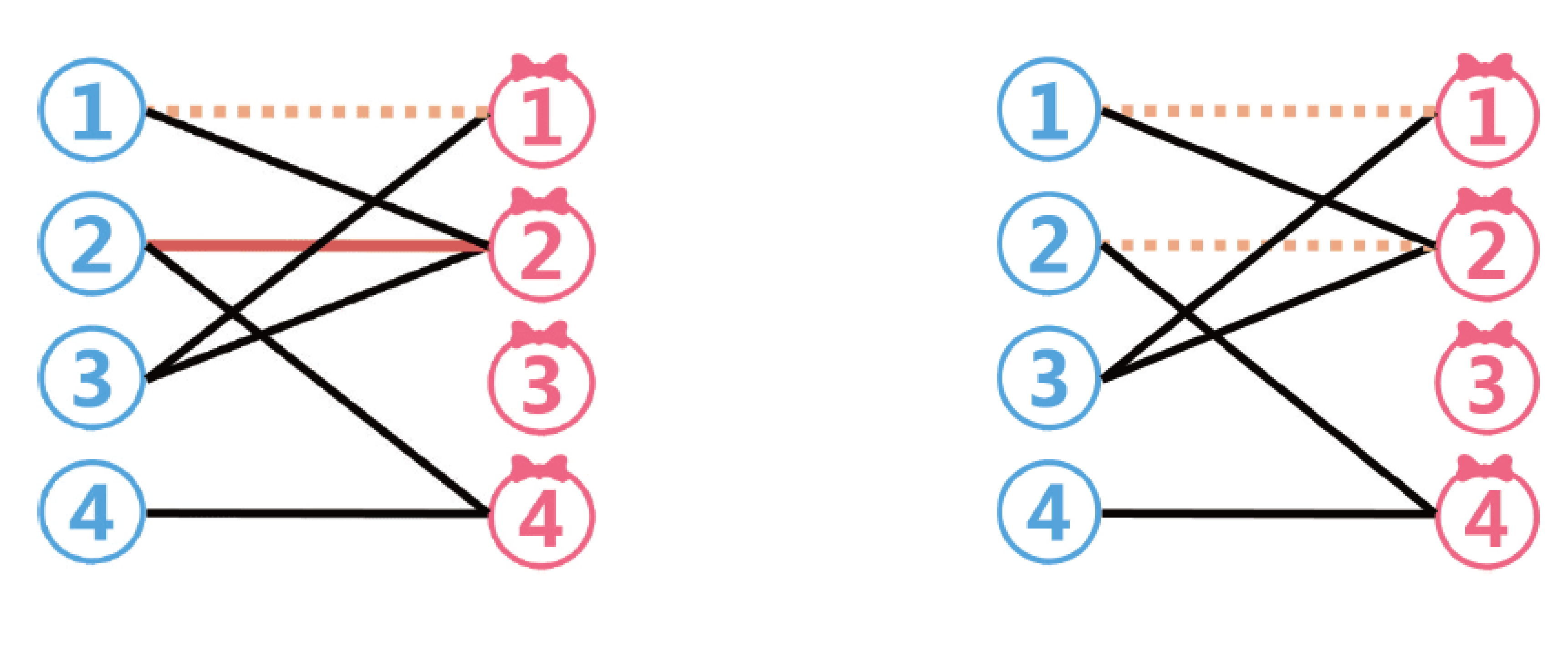


图 3.10　拆散 1 号男士与 1 号女士　　图 3.11　拆散 2 号男士与 2 号女士

此时发现 2 号男士还能找到 4 号女士，那么之前的问题迎刃而解了，回溯回去。2 号男士可以与 4 号女士配对（如图 3.12 所示），1 号男士可以与 2 号女士配对（如图 3.13 所示），3 号男士可以与 1 号女士配对（如图 3.14 所示）。

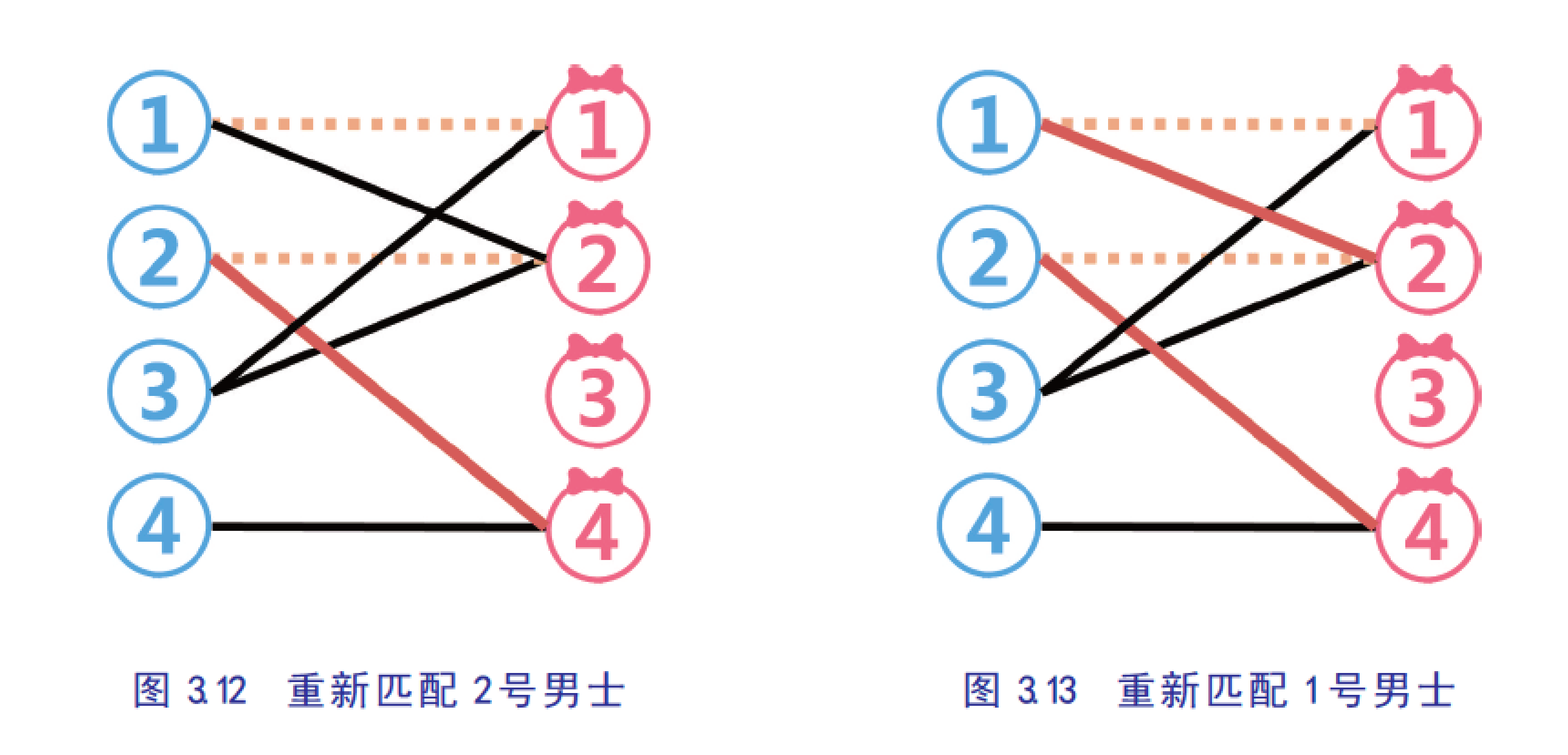


图 3.12　重新匹配 2 号男士　　图 3.13　重新匹配 1 号男士

所以第三步最后的结果如图 3.15 所示。

第四步：接下来是 4 号男生，很遗憾，按照第三步的匹配结果，我们没法给 4 号男生找到一位合适的女士。这就是匈牙利算法的流程，可以看出，其

中找女士是个递归的过程。

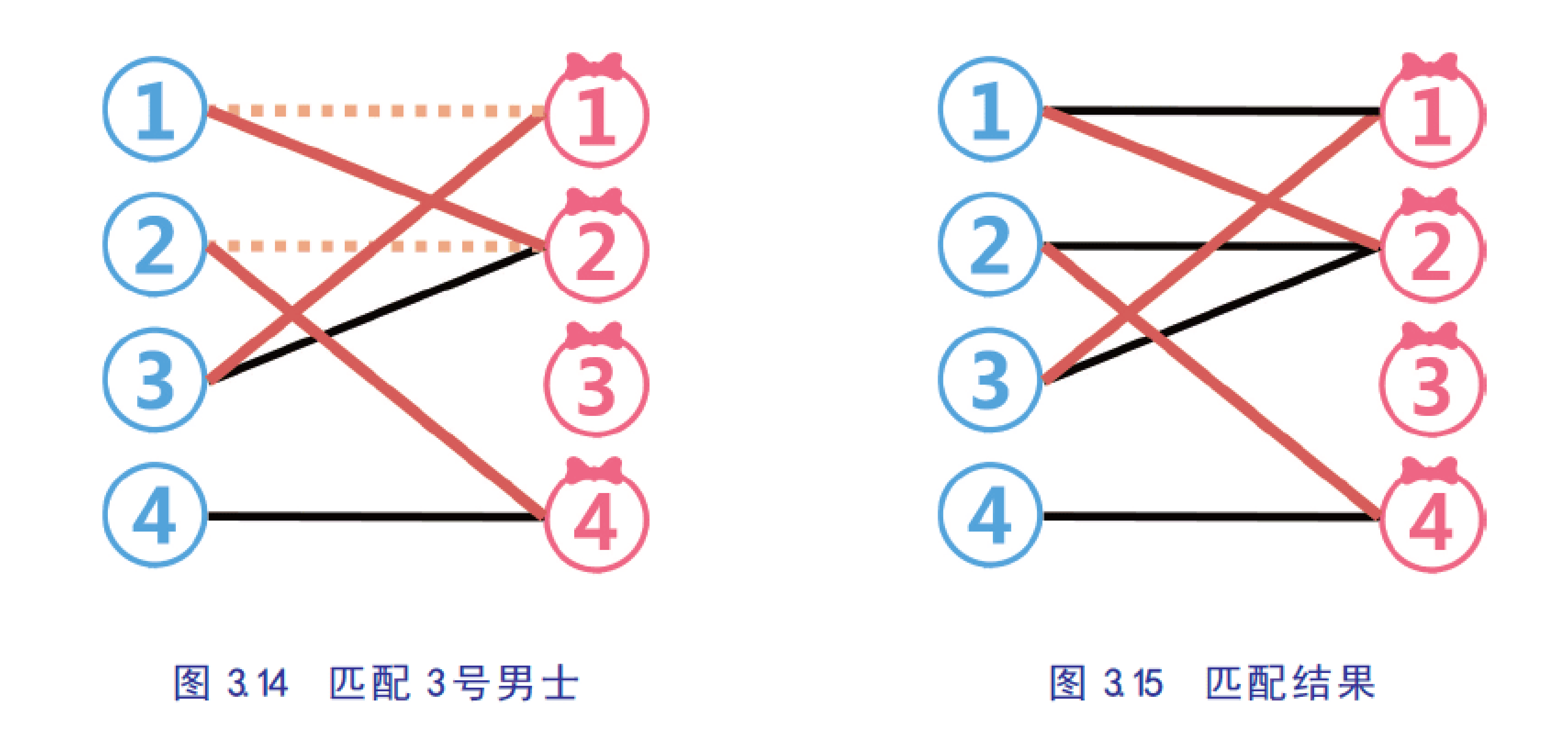


图 3.14　匹配 3号男士　　　　图 3.15　匹配结果

然而,丘比特的目标是找到缘分值之和最大的匹配。也就是说,匹配对象之间带了权值,这称为最大权匹配问题。这里,“丘比特之箭”称为 KM 算法。KM 算法是匈牙利算法的扩展,是解决匹配市场中二分图匹配的另一类经典算法。

KM 算法就是换个角度看二分图的最大匹配,即二分图的每条边的默认权重为 1,所求二分图的最大匹配权重自然是最大的。而对于带权二分图,其边有大于 0 的权重,我们需要找到一组匹配,使全部匹配边的权重之和最大,这就是带权二分图的最佳匹配。

如何从多对一的匹配中获得最佳选择

人生是一场负重的狂奔,需要不停地在每一个岔路口做出选择。而每一个选择,都将通往另一条截然不同的命运之路。

——沧月,当代奇幻文学作家

在婚姻匹配中，我们考虑的男女双方匹配结果是一对一的（一夫一妻的婚姻制度）。但有时匹配结果是多对一的，比如高考招生和公司招聘，多个学生可以分配到同一所学校就读，多个求职者也可能申请到同一个公司就职。

多对一的关系，听上去对考生和求职者很不利。但事实上，如前面所说，这是一个双向匹配市场，只有学生和求职者先做出选择，才会被选择。这样的多对一双向匹配市场情况比一对一更为复杂，如果把握不好，很可能得到最差选项。例如，在高考后，我们经常会发现“高分低就”的现象，即考了高分却只收到低录取分数院校的录取通知书。打个比方说，小高是 2013 理科考生，高考总分 523 分，超过一本线 74 分，但是他却只收到了最后一个志愿南华大学的录取通知（如表 5.1 所示）。

表 5.1　小高的志愿填报表

平行志愿	报考院校	2013 年录取最低分	报考策略
第 1 志愿	华南理工大学	563	冲刺
第 2 志愿	兰州大学	526	稳当
第 3 志愿	首都经济贸易大学	525	稳当
第 4 志愿	南方医科大学	524	稳当
第 5 志愿	南华大学	462	保底

之所以产生这个问题，有两个原因：首先，**信息不对称**。在选择院校时，人们往往只是基于往年录取的最低分数线来选择学校。然而，如果情况发生变化，就会使判断产生重大偏差，产生如同上面例子中所看到的现象：第 1 志愿华南理工大学自 2009 年以来录取分数线每年分别超过一本线 73、85、100 和 113 分，而且呈现逐年上涨的趋势，指望该校大幅降分录取很不现实，如果分数不够，只能寄希望于第 2、3、4 志愿。而小高第 2、3、4 志愿选择了比华南理工低一个档次的相同水平院校，按理说这是合适的，但问题在于不能精准地预测院校录取分数，只是凭往年的排名和录取最低分判断，三所院校

均仅仅差几分而遗憾没有被投档。小高用南华大学作为保底志愿,从最终结果来看,虽然确实保证了顺利录取,但是他本来可以去更好一些的学校。其次,**对平行志愿机制下的匹配算法缺乏了解**,即平行志愿"不平行"(后面会分析到)。如果小高的志愿填报能够避免以上两个问题,完全可以"高分高就",进入更好一些的院校,比如广东外语外贸大学(录取最低分 516 分)、深圳大学(录取最低分 518 分)和南京财经大学(录取最低分 519 分)等。

知分不如估分:个体理性下的集体非理性

学生填报志愿是完成学生与高校之间匹配的第一步,然而在中国,填报志愿方式呈现出多样化的特征。自 1977 年恢复高考以来,中国高考志愿填报方式经历了考前填报志愿、考后估分填报、考后知分填报三个阶段。如何更好地将学生与高校匹配一直是一个难题。

最初,考生先根据自己的兴趣和平时学习成绩填报志愿,然后参加高考,高校根据高考成绩进行录取,这就是考前填报志愿。然而,随着考生人数的不断增加,为了提高匹配的合理性,部分省市开始实行考后估分填报志愿,考生先参加高考,根据估算出来的分数填报志愿,高校根据高考成绩进行录取。大约到了 21 世纪初,大家发现高考前和高考后估分填报志愿的方式,反而加大了考生的落榜风险。于是一些省市让考生先参加高考,等分数出来后再填报志愿,高校根据高考成绩进行录取,这就是考后知分填报志愿。这种方式最初仅在个别省市推行,如今已被我国大部分省市采用。长期坚持高考前填志愿的上海市也于 2017 年改为了高考后知分填报志愿。

虽然知分填报志愿有利于考生选择学校,避免估分填报的盲目性,然而在实际操作过程中却常常会出现各批次竞争加剧,反而更容易造成高分考生落榜的现象。这个现象的产生完全符合经济学中著名的悖论"囚徒困境"。

“囚徒困境”是 1950 年美国兰德公司的梅里尔·弗勒德(Merrill Flood)和梅尔文·德雷希尔(Melvin Dresher)提出的,后来由顾问艾伯特·塔克(Albert Tucker)以囚徒方式阐述,并命名为“囚徒困境”。两个共谋犯罪的人被关入监狱并被独立审讯,不能互相沟通情况。检察官向两人宣布如下规则,如果两个人都不揭发对方,由于证据不确定,每个人都将判刑一年;若一人揭发,而另一人沉默,则揭发者因为立功而立即获释,沉默者因不合作将判刑五年;若互相揭发,则因证据确实,二者都判刑两年。由于囚徒无法确定另外一人会采取何种策略,因此倾向于互相揭发,而不是同守沉默。最终结果是两人都判刑两年,而明显更好的两人共同沉默的结果则无法实现。囚徒困境所反映出的深刻问题是,人类的个人理性有时会导致集体的非理性——聪明的人类会因自己的聪明而作茧自缚。

本来,在考前填报志愿和考后估分填志愿时,考生参考平时成绩,填报时往往比较谨慎。当多数人都比较谨慎时,落榜风险一般说来比较小。而对于考后知分填报志愿,考生明确知道了自己的分数和排名,在心理上不确定性降低了。换句话说,比较敢填了。当所有人都比较胆大的时候,自然会“扎堆”到某一所高校,导致这所高校的报考人数超出招生计划数,从而使考生的落榜风险增大了。这就是上面提到的囚徒困境中反映出的问题:人类的个人理性有时会导致集体的非理性。而这种非理性也常常会出现在其他领域中,比如在国内,当某个商业模式获得成功时,往往会出现创业者和投资人扎堆进入该领域的现象,从而让整个行业迅速进入红海,大家的日子都变得不好过起来。

有时候更重要的是如何选择

既然不同的志愿填报方式都有其问题,那么高校录取机制究竟还有没

有改善的空间呢？虽然 Gale-Shapley 匹配理论已经日趋成熟，但高校录取机制似乎还有很多值得探索的地方。在计算机还没有普及使用的年代，国内外经常使用的录取机制都停留在最为简便的波士顿算法。原因是，过于复杂的机制算法不但使教育主管机构无法承受，而且更容易在操作过程中产生错误。

波士顿算法（Boston Algorithm）在很长一段时间内被美国的公立学校广泛采用。在这个算法中，每个学生对所有 N 个学校进行偏好排序并作为录取志愿提交后，录取规则如下：

第 1 步：每个学校考虑把本校排在第 1 志愿的学生，并把这些学生排序。如果录取名额大于等于学生数，录取全部学生；如果录取名额小于学生数，则录取排名靠前的等于录取名额数的学生，退回多余学生。

……

第 k 步：每个学校考虑把本校排在第 k 志愿的学生，并把这些学生排序。如果剩余录取名额大于等于学生数，录取全部学生；如果录取名额小于学生数，则录取排名靠前的等于剩余录取名额数的学生，退回多余学生。

直到完成第 N 步，或者全体学生被录取，或者全部学校名额招满。

我国过去的高考录取机制与波士顿算法基本类似，但实践表明，波士顿算法是不稳定的，在该算法下，“高分低就”现象非常普遍。

2003 年，美国纽约地区的公立学校开始将 Gale-Shapley 匹配理论运用到学校录取工作中，就连波士顿的教育主管部门也在 2005 年放弃了著名的波士顿算法，开始采用新的招生方式。Gale-Shapley 匹配理论在高校录取机制中的应用，称为“延迟接受”算法。

延迟接受（Deferred Acceptance Algorithm）算法，即每个学生对所有 N 个学校进行偏好排序并作为录取志愿提交后，按照如下规则录取：

第 1 步：每个学校考虑把本校排在第 1 志愿的学生，并把这些学生排

序。留下最多为录取名额数量的学生进入保留名单，退回其他学生。

……

第 k 步：对于上一步被退回的学生，他们的名字进入下一个志愿学校的考虑名单。学校对于新进入的学生和保留名单里的学生进行统一排序，留下最多为录取名额数量的学生进入保留名单，退回其他学生。

直到全部学生的志愿都被考虑过一次或者全部学生进入保留名单为止，此时保留名单即为最终录取名单。

理论上说，延迟接受算法匹配结果是稳定的，但这个的前提是学生和高校可选的范围没有受到“外在因素”的限制。然而，面对近千的高校和数万的专业选择，不可能要求每个考生进行完整排序。在这里，“外在因素”就是：只能允许学生提交有限的志愿数。由于可供学生排序的学校数目远远小于高校总数，因此，学生的选择是非常受限的。

选择是匹配的第一步，也是最后一步。俗话说，“七分成绩定，三分志愿拼”。有时候我们不得到不无奈地发现，“选择比努力重要”，选择不同的大学，人生轨迹或许将会截然不同，而且，对大部分人来说，只有一次选择的机会。从某种程度上，可以说不同的匹配算法，决定了一个人不同的人生走向。

中国式匹配：平行志愿“不平行”

波士顿算法和延迟接受算法都允许每个学校自己制定排序方法，而不只是根据学生的考分。而近些年教育部力推的“**平行志愿**”**更加类似于分数独裁机制**。对于所有学生来说，每个学校的排序依据都是相同的，即考试分数。同样，在这里，我们不可能要求每个学生对上千高校和数十万专业进行完整排序，只能允许其提交有限数量的平行志愿。

在平行志愿录取机制中，每个学生对所有 N 个学校进行偏好排序并作

为录取志愿提交后，录取规则如下：

第 1 步：分数排名第 1 的学生，第 1 志愿学校立即录取该生。

……

第 k 步：分数排第 k 名的学生，顺序考虑他的第 1、2、……、N 个志愿学校，如果某个学校的录取名额还没有满，则该校立即录取该生。

直到全部学生的志愿都被考虑过一次或全部学生都被录取为止。

如果实行了平行志愿，即使你不能被第 1 志愿录取，仍然可以被水平相当的其他高校平行录取，这样就避免了落榜风险。这就是目前多数省市实行的高考后知分填报志愿加平行志愿的方式。

有人用“车站上车”来比喻平行志愿的录取方式。

一个巨大的车站。停车场上停了几百辆汽车，每一辆车代表一个高校，每辆车的座位数量代表了该校在该省的录取人数。其中有的是中巴车，比如南京审计学院，只有 9 个座位。有的是超级大巴车，比如四川大学，座位超过 1000 个。

阳光明媚的早晨，全省考生按成绩排名，站成一路纵队，每个人手里都有一张纸条，上面写有自己的 5 个平行志愿的学校名字（A、B、C、D、E）。

首先，车站入口打开，状元先进，然后入口关闭。此时，偌大的停车场，只有状元一人。状元的 A 志愿是清华大学，于是走到清华大学的中巴车前，座位全空，顺利上车（**投档**）；上车后手里的小纸条作废，后面的 B、C、D、E 志愿再无任何用处。状元上完车了，轮到榜眼了，同样只有他一个人进入停车场，此时探花还需要等一等。榜眼的 A 志愿是北京大学，同样座位全空，于是榜眼顺利上车，其纸条作废。

按顺序来，比如，该排 200 名的同学进场了，A 志愿清华大学，到清华大学的中巴门口一看，80 人的位置已经坐满了；转头又看 B 志愿北京大学，一看 70 人的位置也满了；再看 C 志愿浙江大学，50 人的座位只坐了 6 人，于是

该同学上车,其纸条作废。

现在轮到排 1000 名的同学进场了,A 志愿北京大学,满员;B 志愿复旦大学,满员;C 志愿武汉大学,满员;D 志愿四川大学,也满员了,此时该同学开始冒汗;赶快看 E 志愿青海大学,到青海大学的大巴门口一看,还好,300 个位置还有 150 个空位,赶紧上车,其纸条作废。

现在轮到排 5000 名的同学进场。A 志愿西南大学,B 志愿厦门大学,C 志愿湖南大学,D 志愿华中师大,E 志愿华南师大,5 辆车全部都满员了,长歌当哭! 转头看到旁边的天津大学还有 2 个空位,想强行上车,被车站保安制止,纸条遭没收,被赶出停车场,悲剧收场。

该最后一个学生了,排 26 000 名进场。此时大多数汽车已满员,个别车还有空位(因为前面有 1000 个不幸被赶出车站的同学)。赶紧把纸条摸出来看: A 志愿川农,B 志愿川师,C 志愿北京工商,D 志愿华南农大,时运不济,所有车全满员。再看 E 志愿天津大学,还有 2 个位置,谢天谢地! 上车走人。

所有人员全部处理完毕。车站站长一声令下,几百辆汽车全部开出车站(**所有高校是同时投档的**)。省招办的工作暂时告一段落。

四川大学的大巴车上。售票员进行座位分配(**选专业**),如果分数不够高,选不到好位置(**每校可填 6 个专业志愿**),又不同意售票员安排座位的(**不同意专业调配**),那只好请你下车了(**退档**)。于是灰溜溜地回到汽车总站,与刚才被赶出车站的 1000 多人站在一起,此时停车场上空无一车。

时间到了下午。有 30 辆当初没有收满的车又开回来了(**征集志愿开始了**),1000 个当初被赶出车站的,以及 500 个被售票员赶下车的同学,大家重新排队,重复早上的过程。

那么,为什么第 5000 名的同学的纸条会遭没收,并被赶出车站呢?

因为该同学高估了自己,写纸条时又不认真,犯了低级错误: B、C 和 D

志愿没有任何意义(白白浪费了三个名额!)。因为,即使该同学是全国状元,他也上不了厦门大学、湖南大学或华中师大的巴士,因为前面A巴士(西南大学)一定把人“截走”。

有人问:“我不上A巴士,上其他更好的不可以吗?”

答案是:“不可以,必须按A、B、C、D、E的顺序来!”

“什么,平行志愿怎么‘不平行’了?”

是的,这才是问题的关键。对考生个人来说,平行志愿本来就是不平行的!

如果该同学的E志愿不是填华南师大,而是广西大学(还有5000个空位),那么他就可以上广西大学这辆超级大巴了。悔不当初!

平行志愿不平行。但是,与以往的顺序志愿仍然具有较大的差别。

顺序志愿的方式,车站外不用排队,2万多人一起涌入停车场,各人按自己的纸条到对应的汽车前面排队。由于顺序志愿的一个志愿只有一个学校,所以你只能在一辆汽车前面排队。按成绩排好队以后,依次上车,满员为止。例如,第1志愿清华大学,第2志愿四川大学,现在你只能在清华大学的车前面排队,如果你排在清华大学队列的最后一名,估计很可能上不了车,开始冒汗。回头一看,四川大学汽车前的同学的分数都比你低,但你也只能继续排在清华大学的队伍里死等(其实并没有希望),眼睁睁地看着别人上四川大学的车。同样,没有上车的同学当即就被赶出车站。然后等第1轮没有收满人的车开回车站,开始第2志愿的排队。

那么这种机制的效果究竟如何呢?通常针对匹配机制的评价是基于这4项指标:首先是稳定性,由于在匹配中没有价格存在,匹配的稳定性替代了均衡价格和数量的概念,表示匹配的均衡状态,即匹配的最终结果不会发生自愿的重新匹配;其次是无浪费的性质,即匹配机制使得愿意匹配的离散资源尽可能多地完成匹配,也就是常说的“物尽其用”;第三是讲真话的性质,即

参与人不能通过虚报自己的序数偏好以获利;第四是帕累托有效(Pareto Efficient)的性质,即不存在一个新的匹配结果,在不降低其他离散资源的效用的同时使得至少一个离散资源的效用得到提高。

平行志愿与延迟接受机制的重要区别在于,延迟接受机制是以高校为中心的(高校可以充分地对比所有申请者,选出最满意的学生),而平行志愿是以学生为中心的(算法会优先处理分数高的学生)。前面说过,由于学生的选择是有限的,因此,不论是波士顿机制、延迟接受机制还是平行志愿机制,其匹配结果都是不稳定的。即使如此,其不稳定的程度仍然不同。有人通过实验比较了平行志愿、波士顿和延迟接受机制,发现**平行志愿机制的稳定性高于波士顿机制,但是低于延迟接受机制**。另一方面,由于学生只能提交一次高考志愿,对应到算法里,就是学生对高校的偏好列表是一开始就确定的。因此,三种机制都无法有效避免**"大小年"现象**。例如,某高校去年的录取分数非常低,今年可能意外地非常高。"大小年"往往导致学生"扎堆"到某一所高校,导致这所高校的报考人数超出招生计划数,或者某一所高校"无人问津",导致这所高校无法招到满意的学生。

一种可行的方案,可以允许学生多次填报志愿,即学生的偏好可以根据预填结果动态调整。然而,这样的方案背后需要强大的算法和复杂的操作来支撑,会导致成本大大增加。对此,我们只能根据招生中使用的匹配算法规律,给出一些针对平行志愿机制下的填报建议:

高考考试完不是战斗结束而是开始,志愿的填报是第 2 次高考,这时考的不是努力程度,考的是做选择的能力。这时,选择比努力重要。

尽量使多个平行志愿之间具有梯度。有人片面地认为平行志愿没有顺序,几所院校都一样,这是不对的。录取时是按照考生自己填报的顺序依次进行检索的。因此,一定要把最想去的院校放在最前面。第一所院校的录取分数可以高一些,往后要按照院校今年可能的录取分数梯次降序排列。否

则，若第一所院校录取不上，后面几所院校有可能都录取不上，成为无效志愿。

估分填报时，在理性估分后多参考历史经验，将"高分低就"风险降到最低。

知分填报时，热门或者冷门的高校和专业要谨慎选择。因为平行志愿机制下仍然没有避免出现"大小年"现象，所以要科学地分析院校录取数据。

互联网思维≠没有中间商赚差价

在市场上常常可以看到一种情况：那个叫喊得最凶的和誓发得最厉害的人，正是希望把最坏的货物推销出去的人。

——列宁

志愿填报的例子也说明了匹配市场的信息对称性问题。在知分填报志愿的方式下，大部分人都不会仅仅以自己和周围同学的分数作参考，而是想方设法去了解全省报考某一所高校的考生的数量，并认为只有这样才能做出更加准确的判断。了解全省几十万考生信息与了解全校信息相比，其成本显然不是一个数量级。而最终能否被录取，则完全取决于全国到底有多少人填报了这所高校的这个专业。而考前和考后估分填报志愿，考生搜集信息花费较少的成本，通过往届考生成绩、排名、录取院校情况即可做出判断。因此，不同的高考志愿填报方式下，考生所面临的信息搜集成本是不同的。**信息搜集成本的存在，导致了因个体差异引发的信息不对称。**

在匹配中，信息的对称性会影响匹配的结果，掌握更多信息的一方在经

济活动中往往处于更有利的位置。玩过游戏的朋友可能都有这个经历：当你的地图上全是战争迷雾，而对方地图清晰可见时，你很难赢得游戏。《史记》里“田忌赛马”讲的也是一个利用信息不对称而获胜的故事。田忌掌握了所有马的速度和齐威王出马的顺序，这有利于他更好地制定比赛策略。

齐威王要与田忌赛马，规定每个人从自己的上、中、下三等马中各选一匹来赛；每有一匹马来比赛；每有一匹马取胜可获一千两黄金，每有一匹马落后要付一千两黄金。齐威王的每一等次的马比田忌同样等次的马都要强，因而，如果田忌用自己的上等马与齐王的上等马比，用自己的中等马与齐王的中等马比，用自己的下等马与齐威王的下等马比，则田忌要输三次，要输黄金三千两。田忌的谋士孙膑让田忌用自己的下等马去与齐威王的上等马比，用自己的上等马与齐威王的中等马比，用自己的中等马与齐威王的下等马比。田忌的下等马当然会输，但是上等马和中等马都赢了。“三局两胜”，田忌不仅没有输掉黄金三千两，还赢了黄金一千两。

信息不对称让匹配市场难以用最高效率实现，令供求双方蒙受损失。

20 世纪 70 年代，美国经济学家乔治·亚瑟·阿克洛夫曾发表过一篇名为《柠檬市场：质化的不确定性和市场机制》的论文。在二手车市场中，买主无法辨别汽车高低质量的区别，只有卖主才知道这一点，因而在可获得的信息方面出现了不对称性。低质量汽车与高质量汽车以完全相同的价格销售。由于价格空间浮动较大，最终低质量汽车很可能将高质量汽车驱逐出市场。这在经济学中称为“劣币驱逐良币”。信息不对称问题可能导致整个市场崩溃，或者市场萎缩，以至于只有劣等产品充斥其中。2001 年，瑞典皇家科学院将诺贝尔经济学奖授予乔治·亚瑟·阿克洛夫等三位经济学家，以表彰他们在市场信息不对称研究中所做出的卓越贡献。

信息不对称导致的匹配问题因互联网的兴起得以改善。利用互联网技术并发挥匹配作用的各个平台也应运而生，在租房中介、二手车、二手房交易

等市场都产生了相应的创新型商业模式。近年兴起的“互联网思维”将“互联网消除信息不对称”的能力，理解为“凡是因信息不对称而获利的公司都会被互联网颠覆！”。然而，事实真的是这样吗？

假如没有中间商

酒桌常有“行家”告诫大家说：“一瓶酒 300 块，120 块是广告费，经销商拿了 100，厂家自己挣 50，酒的成本其实只有 30 块”。听“行家”说完，大部分人可能都会产生“买贵了”的感觉。

然而，如果没有经销商和广告商，那么消费者与酒厂之间的匹配渠道将由谁来完成呢？事实上，在复杂的市场环境中，由于中间商的存在，交易环节不增反减，交易成本不增反降。如果削掉合理的中间环节，消费者拿到手里的商品价格只会升高而不会降低。正是因为**中间商“知道得太多了”，因而简化了交易关系，减少了交易双方为了达成交易所需联系的数量和时间，让交易成本更低。**

如图 3.16 所示，A 部分显示了三个生产者以直销的方式分别接触三个消费者，在这个系统中，存在 9 次交易联系；B 部分显示了三个生产者通过一个中间商与三个顾客连接，这个系统只要求 6 次交易联系。可以看出，中间

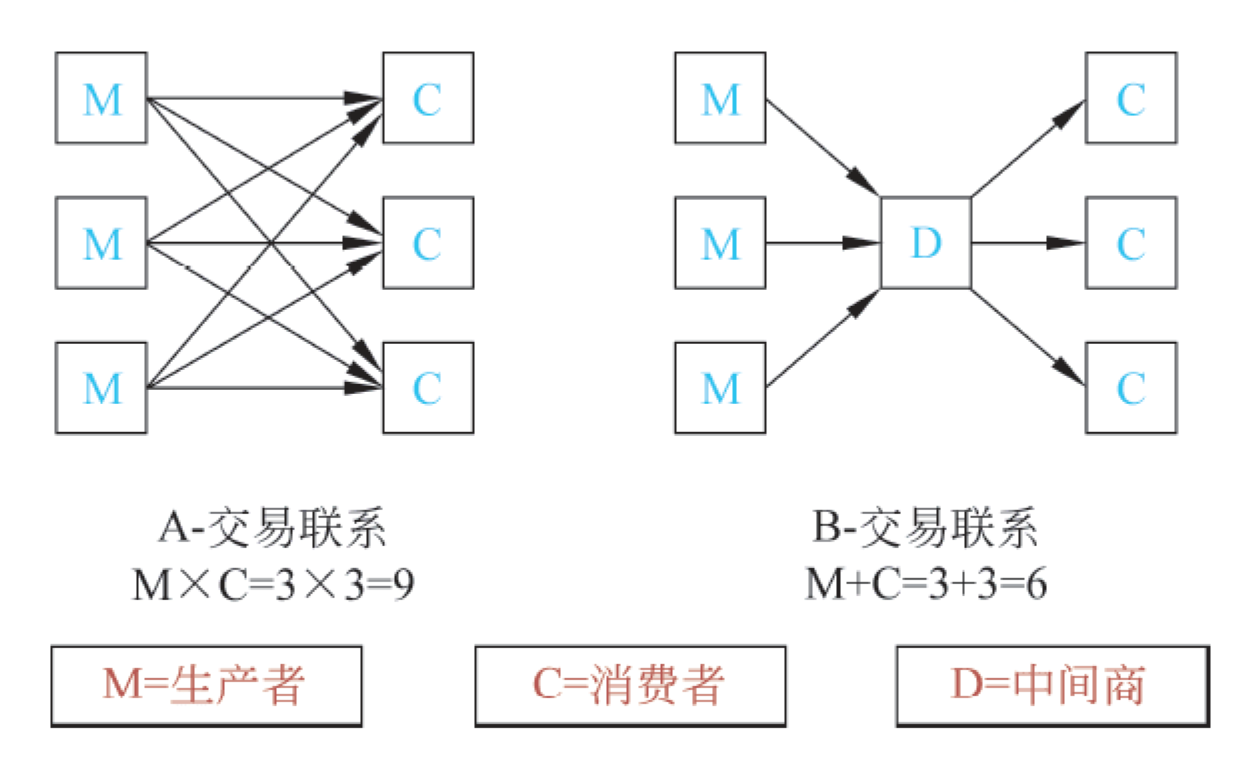


图 3.16 中间商效果图

商的存在减少了必须进行的工作量，而且消费者越多，生产者越多，中间商的价值就越大。

“厂家直销，没有中间商赚差价”和“0 佣金 0 手续费！”等听起来很美好，然而，这一切只是依靠营销手段为消费者编织的童话，并不是真正意义上的没有中间商。其本质是减少中间商参与的交易，而中间商则有了新含义。

匹配，让中间商经济实现向共享经济的进化

“共享经济会在 21 世纪下半叶成为社会主导的经济形态”，未来学家与经济学家杰里米·里夫金（Jerermy Rifkin）的这一预言正在被一步步实现。在赛博新经济的范畴，共享经济当属时下最流行的新经济范式。

将共享经济理念践行得最为成功的，非 Airbnb 莫属。在这一经济活动中，Airbnb 要做好房东与房客之间的匹配，除了根据房东和房客的偏好做个性化推荐以外，还需要解决一个更为棘手的难题，那就是信任体系。

总部位于旧金山的 Airbnb 是短租行业的始祖，自 2008 年上线，它创建了帮助人们将空房和沙发变现的模式，截至目前，Airbnb 在全球 192 个国家 3.3 万个城市总房源超过 230 万，并拥有超过 6 亿个社会关系。

为信任而设计是传统酒店行业最重要的主题，而 Airbnb 的使命是赋予个体最大的自由度，因而会更大程度地依赖这种信任关系。让一位完全陌生的人住在自己家里，Airbnb 这种想法实在太过疯狂了。毕竟从小妈妈就告诉我们：“不要和陌生人说话”。

所有通过 Airbnb 的交易都可以查到交易发生时使用的信用卡，接着是根据银行办卡时的实名信息，总能追查到某个人。所以主人或房客无论哪一个做坏事，风险都非常之高。在解决这个问题上，Airbnb 利用在线社交网络脸书（Facebook）来帮助其解决这个问题。脸书是很注重实名制的社交平台。

当房客用脸书账号登录 Airbnb，并搜索一个城市的房间信息时，Airbnb 会显示房客和房东之间的社会关系。Airbnb 甚至将 Facebook 的这种社交网络功能推向极致，它做了一个名为“Airbnb 社交关联”的过滤功能，专门给用户提供房客对社交圈内房东的查找。比如，搜出来的房东或者是朋友的朋友，或者是同一个大学毕业的，或者是朋友在那里住过并且写过评价的，这相当于在熟人圈里找住处。通过这个功能搜索房源，搜索获得的所有房东都与房客具有某种社会关系，房客的安全感会提升，对房客来说比住在陌生人家里有更大吸引力。

除 Facebook 之外，如果这个房东的 LinkedIn 被绑定的话，他的可信度也会大大增加。因为与 Facebook 一样，LinkedIn 也是实名制，而且信息包涵用户教育经历、工作经历，以及现在就职的地方，因此，LinkedIn 对美国人找工作已经越来越重要，如果这个账号被盗，将是一件非常严重的事。所以大部分用户都会认为：如果房东愿意把 LinkedIn 绑定到 Airbnb 上，房东似乎在传达一个意思：我做 Airbnb 是认真的，以工作为保证。另一个评估房东可信度的重要方式是，查看房东收到的以往房客对他的评价，如果以往房客都给他好评的话，就可以大概判断该房东是可以信赖的。最后，Airbnb 推出了“身份验证”项目，所有 Airbnb 用户都必须验证身份。

共享经济的终级奥妙就是，**我暂时不用的东西你正好用得上**。租车界的 Airbnb 当属优步（Uber）。作为一个提供私家车搭乘服务的 O2O 网站，它通过 GPS 追踪定位私家车，用户可以使用 Uber 发出打车请求；几分钟内一辆私家车就会开到你面前，费用则是通过信用卡交易来完成。把 Airbnb 的概念用到办公室上就有了 ShareDesk 之类的服务。一方面公司可以提供闲置的办公桌、会议室等，另一方面移动办公族可以在任何地点工作。共享经济出现在越来越多的领域，如 3D 打印机、传授技能和零工等。乌马尔·哈克在《新资本主义宣言》中预言：“如果传统消费减少 10%，而共享消费增加

10%，那么传统企业的利润率将受到显著影响；如果传统企业不能进行改变，甚至可能会被淘汰。”共享经济的出现将在很多领域抢占传统经济的市场。

在共享经济里，中间商演化成了平台，匹配渠道平台化。中间商改了个名字，称为平台商。例如，Airbnb 不仅向租客收取 6%～12%的费用，还会收取房东 3%的附加费用。有些中间商则演化成了纯粹的平台提供者，不再收取任何服务费用，而是利用多边市场寻求其他商业模式，如流量变现、数据变现等。例如，Facebook 的 95%以上收入来自广告，阿里巴巴的主要收入也来自广告和服务费。

“共享”二字蕴含着无限大的想象空间。首先，共享经济让传统的中间商平台化，平台化的匹配行为，让消费者拥有更大的掌握权，交易完全透明化。从直觉上理解，就是参与人知道的信息越多，他的选择就会更优。其次，共享经济中存在很多去中介的交易，有中介的交易，由中介提供担保，而去中介的交易，则需要像 Airbnb 一样设计相应的信用机制，保证匹配的有效性。类似的还有淘宝网，如果买家和卖家的信用分高，匹配就容易完成，因为信用分代表“公民素质”，是影响双向选择的偏好性指标。可以说，**“共享经济时代”是“中间商时代”的进化**，但不论怎么演化，其核心都是，**让信息对称，提高信息匹配效率**。

匹配算法设计下的赛博新经济

在现实生活中，还有很多市场都处于“商品市场”和“匹配市场”之间，市场是可以被设计的。罗斯开创了一个全新的经济学分支——市场设计，即寻求在市场失灵、无法依靠价格这单一因素保证其正常运行时的资源配置方案。罗斯的匹配理论实践，有效改变了公立学校、肾脏捐献以及住院医师岗位配置的市场设计和功能。在共享经济模式中，利用延迟接受算法，为市场

的中心调度制定规则，可以确保选择双方形成稳定匹配，从而优化市场资源配置，提高市场运行的效率。

“钱不是万能的”，因为匹配市场还需要匹配算法这只“无形的手”。“没钱是万万不能的”，因为市场设计就是综合“钱”这一因素的供需匹配机制设计，是研究如何利用“钱”和匹配理论来实现资源最优配置的新学科。**介于“商品市场”和“匹配市场”之间的市场中，最典型的就是拍卖**。拍卖设计是市场设计中最活跃的部分，拍卖有多种组织方式，但不论是什么方式，拍卖最古老的功能之一就是价格发现：市场会告诉你能从所卖商品中寻找到怎样的价位，并且你能够对谁以这样的价格出售。正如罗斯所说：“拍卖是将卖方和最珍视所卖的商品的买方匹配起来的市场”。匹配是一个非常重要的市场设计手段，**匹配算法设计下的市场，也将成为一个更机智、更敏捷的智能经济市场**。

第4章　动态定价，应对供需剧烈变化的赛博新市场

2017年4月9日发生了一件令美联航陷入空前危机的事件。

当天晚上，美联航UA3411航班正准备从芝加哥飞往肯塔基，但乘客登机后，却被机组以机票超售为由，要求其中4名乘客必须下飞机改乘其他航班。为了让乘客主动下飞机，机组人员提出赠送旅行代金券，但这并没有引起任何回应。最后航空公司以随机抽取乘客的模式，请一位69岁的越南裔美国人陶大卫改签，但被他以次日需要出诊为由拒绝。没想到的是，最后这位乘客被机场保安强行拖下飞机，在拖拽过程中，他嘴角磕破出血，衣服也被撕扯……当时在场的其他乘客都被这个场景震惊，拍下的视频随后在全球互联网迅速传播，给美联航造成了严重的负面影响。受超售事件影响，美联航股价于4月11日大跌(如图4.1所示)。

这里我们先不讨论美联航自身的客户关系管理以及此次事件处理的合法性问题，我们想知道的是，为什么航空公司的售票系统发展至今，仍然会有



图 4.1　美联航股价于 4 月 11 日大跌

超售问题发生？为了保护消费者，政府为什么不干脆明令禁止超售？

从 1978 年航空业放开管制到 2005 年，美国已经有 160 多家航空公司破产，兼并收购屡见不鲜；在 1990—1999 年这段时间，航空业的利润率是全美国所有行业平均水平的六分之一；2001 年“9 · 11”恐怖袭击事件的发生，让大多航空公司在之后一段时间处于严重亏损。航空公司之所以越来越难赚钱，与需求一直在动态变化有很大关系。对每个航班来说，供给的数量是固定的，因为飞机上的座位是固定的。对每个航班而言，需求又是复杂多变的，乘客人数取决于很多不确定因素，例如天气、重大活动、经济景气程度等。同时，每个乘客的支付意愿不同，乘客的偏好和需求还可能随时间而变化。所以，既要把给定的供给销售给动态变化的需求方，而且还要最大化收益是非常困难的。而且，航空公司之间的竞争也在不断加剧。乘客需要的是从 A 地用最快的速度赶到 B 地，就这一点来说，航空公司之间提供的服务很难有本质区别。由于产品和服务不容易差异化，价格战就难以避免，“宁可亏钱也不空位”，有些公司甚至把票价降到几乎为零来多吸引乘客。

这也导致了机票超售成为航空公司优化收益的最好做法。机票超售的意思是，航空公司会卖出比飞机上座位数量更多的机票，因为统计数据表明，总会有部分乘客临时取消行程或者干脆到时不出现，这时如果有一定比例的超售，那么飞机上的座位不至于有太多的浪费。图 4.2 给出了美国各大航空公司的机票超售比率，可以看出，机票超售是各大航空公司普遍使用的策略。

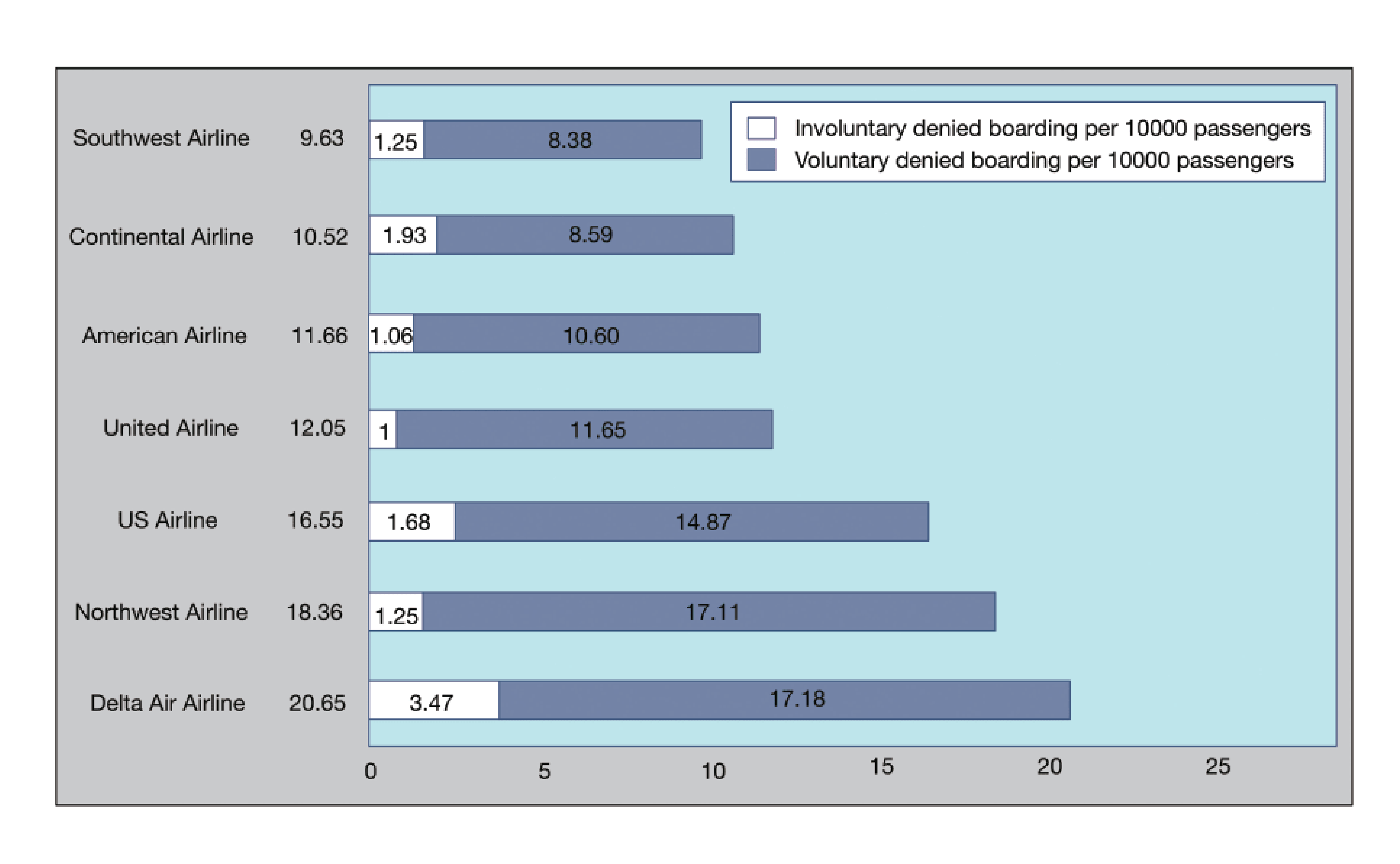


图 42　美国各大航空公司的机票超售比率

服务业的产能是有时效性的，飞机起飞后的空座位，夜晚降临后的酒店空房都是纯粹的浪费，对企业是不可弥补的损失。当今世界的各大航空公司都会超售，也就是卖出比座位数更多的票。超售多少的决策取决于历史数据和不同公司的战略考量。超售最理想的结果就是不出现的乘客数正好等于超售的数量，这样飞机刚好坐满，皆大欢喜。如果临时不来的乘客大于超售数量，飞机上会有空位，航空公司将承受机会损失。如果临时不来的乘客小于超售数量，肯定会有人上不了飞机，航空公司必须提供补偿，寻求自愿换到下个航班的旅客；如果提供了很高的补偿数额还是没有足够多的志愿者，航空公司就会像前面提到的美联航事件一样，不得不对乘客下逐客令了。

美联航的超售事情，也让我们发现超售是有风险的，它有可能给消费者带来极坏的用户体验，还可以导致航空公司的信任危机，但是政府部门却一般不会明令禁止航空公司这样做。如果我们换个角度看，这样做也并没有什么问题。首先，归根结底，超售对大多数消费者有好处。如果不允许超售，那么航空公司空座率会更高，为了盈利，航空公司会提高票价，最终还是消费者买单。其次，超售对社会福利是有好处的。从社会整体来看，我们需要最有效地利用资源，同时避免浪费资源。因此最优策略是尽量让每架飞机都坐满，否则每个空座位都是对资源的浪费。

另一方面，“市场决定价格”的市场经济在全球发展一百余年，在中国也已经开展了二十几年，更多时候市场的力量比法律的手段更有效。政府需要做的就是让市场干净公平，剩下的事情，市场的最终选择会给出一个最正义的结果。而之所以市场会这么强大，只因为经济学有一个最基本的假设，即“理性人假设”。当我们讨论理性人时，我们所有的市场行为和决策都会从最有利于自己并让自己利益最大化这一点出发，而不考虑社会伦理道德等其他因素。例如，在 2005 年，五级飓风卡特里娜席卷美国港口城市新奥尔良，随之而来的巨浪酿成洪水惨剧，由于没有足够的物资储备，发生了哄抬价格事件，药物食品抢售一空。人们也许会抨击在灾难之后提高物价的店家不人道，但是按照理性人的假设，当面对明显增加的需求时，供给的不足自然会导致价格的急剧上升，这其实非常合理。因此，在市场机制比较完善的国家并没有《价格法》，只有《反垄断法》。

当然，遇到超售改签并不是一个大概率事件。根据数据统计，大约每 1000 人中有一个人会被超售影响，自愿或非自愿地更换航班，被强制改签的乘客（非自愿更换航班）大约是每万人中有一到两个。所以美联航事件虽然会让超售在一段时间内成为争议的焦点，但可以预料的是，航空公司并不会因此改变超售的策略。

除了超售，航空公司还有一个常见的策略是机票打折。大家可能都有体会，如果行程能够尽早确定，那么早订机票通常会更加便宜。随着出行日期的日益临近，机票价格则会不断上涨，直到机票售完。当然还有一种情况，如果直到飞机临起飞前，机票仍然没有售完，可能会以很低的价格出售，如果你冒着不能成行的风险等到了最后一分钟，可能会遇到很大的价格惊喜。这个价格变动的过程其实就是动态定价。这里描述的是简单情况，其实航空公司的机票价格是一个很复杂的过程，本章后面会有更详细的介绍。

实际上，超售问题体现了供需的矛盾，这种矛盾随着赛博新经济时代的到来，在一些新的商业模式中表现得更为突出。例如在出行领域，像滴滴出行这类公司，客户的需求变化更为剧烈，对价格调节供需的实时性要求也就越来越苛刻，响应时间从每天到每秒，这背后依靠的就是不断优化的动态定价算法。可以说，航空公司是动态定价算法的早期实践者，通过动态调整票价把有限资源做动态分配，虽然目前还不能完全避免超售带来的问题，但是它为动态算法的发展和演进提供了思路。动态定价算法的进步也在不断地推动越来越多的行业发生商业模式的变革。

动态定价不仅仅是一系列传统定价方法的策略组合

什么是“精明”的人？

他通晓世间万物的价码，

但对其价值却一无所知。

——奥斯卡・王尔德，英国作家

也许大家都有“花冤枉钱”的经历，刚刚在实体店买的正价商品，回到

家上网一看,呀,网上商城正在打折出售,或者才在网上下单的电子产品,几天时间价格竟下跌了一半。是的,商家总会为了卖出更多的商品,根据渠道、客户、商品的不同,在一定时间内打折促销,这就是最简单的动态定价。

在传统的市场定价中,商品往往先定价再出售。在制定合理的价格时,“产品生命周期理论”(product life cycle,PLC)常被用到。这是美国哈佛大学教授雷蒙德·费农于 1996 年提出的。费农认为:产品和人的生命一样经历引入、成长、成熟、衰退 4 个阶段(如图 4.3 所示),处于不同产品生命周期的人群有不同的消费特点。

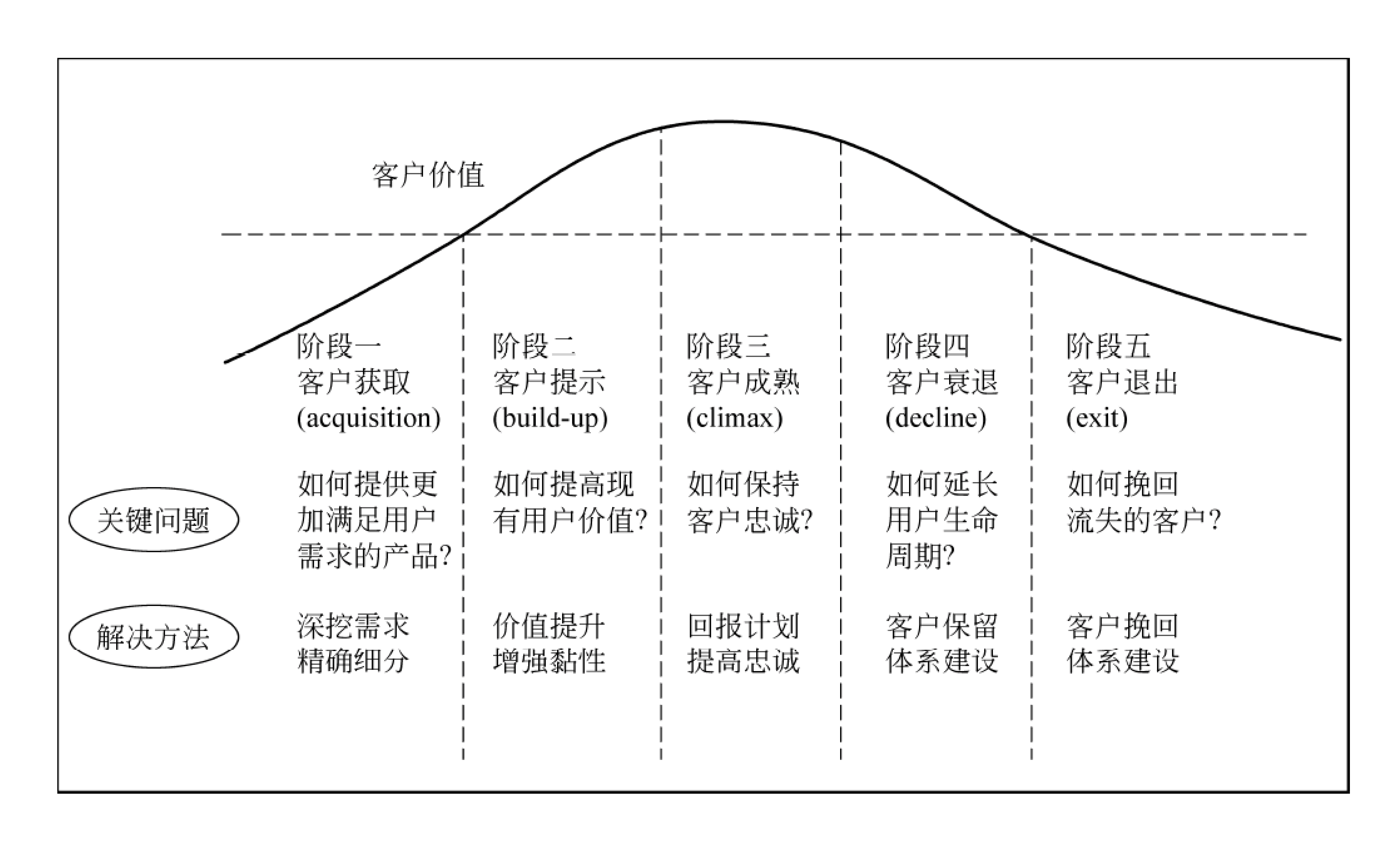


图 43　产品生命周期图

引入期:产品从设计投产到投入市场进入测试阶段。产品种类少,顾客对产品不够了解,除少数猎奇心理的顾客外,几乎无人大量购买该产品。生产者为扩大销路,需要投入大量的促销费用,对产品进行宣传推广,而由于生产技术的限制,产品生产批量小成本高,导致销售价格偏高。

成长期：产品通过试销效果良好，购买者逐渐接受该产品，进入需求增长阶段。此时，竞争者看到有利可图，纷纷进入市场参与竞争，使同类产品供给量增加，价格随之下降。

成熟期：产品走入大批量生产并稳定进入市场销售，购买人数增多，市场需求趋于饱和。

衰退期：产品进入淘汰阶段，随着消费习惯的改变，产品销售量和利润持续下降，市场逐渐出现其他性能更好、价格更低的新产品，以满足消费者新的需求。旧产品无利可图，撤出市场。

根据产品所处生命周期、产品的市场竞争状况、产品的成本结构等因素，传统的市场营销学有几种经典的定价方法，即成本导向定价法、需求导向定价法、竞争定价法等。

成本导向定价法

商品定价的第一步，一般是计算成本，对于普通的大众消费品，先计算产品的原材料成本，再加上广告推广成本，以及人工、房租、水电等运营成本，得到单个产品的成本。成本导向定价法由分为成本加成、边际成本、目标利润三种。

（1）成本加成：在总成本的基础上，加上一定百分比的加成，制定出产品的销售价格。

（2）边际成本：企业定价时，只考虑变动成本，抛开固定成本，而以预期的边际贡献适当补偿固定成本。边际贡献是指预期的销售收入减去变动成本后的收益。

（3）目标利润：根据企业的总成本和估计的总销售量，确定一个目标收益率，作为定价的标准。

其中，第（1）种方法中渠道商和零售商用得较多，另外两种品牌商或工厂什么的都有人使用。例如，出版社出版图书的成本就非常复杂，固定成本有

初稿的排版、制表、制图等，印刷成本有印刷、纸张、装订等，此外，还有营销、设计、库存等间接成本，所以，大部分出版社图书的定价往往采用生产成本的加成方法。

需求导向定价法

企业在定价时，不再以成本为基础，而以消费者对产品价值的理解和需求强度为依据。需求定价又可以基于理解价值的定价和基于需求差异的定价。

（1）理解价值：以消费者对商品价值的感受和理解程度作为定价的基本依据。把买方的价值判断与卖方的成本费用相比较，定价时更应侧重考虑前者。

（2）需求差异：以不同时间、地点、商品以及不同消费者的消费需求强度差异为定价的基本依据，针对每种差异决定其在基础价格上增加还是减少价格。

我国的快递行业目前竞争非常激烈，在电商异军突起的几年中，伴随大量订单产生了剧增的物流需求，相应地，整个行业的基础定价也水涨船高。由浙江杭州桐庐县人创立的“四通一达”（申通快递、圆通速递、中通快递、百世汇通、韵达快递）在2014年11月开始涨价，平均涨幅1.5～2元/单，接着，几家公司在官网上发出声明，声称此次涨价原因是燃油价格上调和人工成本增加。实际上，油价和人工费只是涨价的诱因，真正的驱动力是市场需求的变化。中国快递业务量增长有80％来自电商，而快递是连接电商和消费者的最后一公里，由于服务同质化严重，在需求增长的最初阶段，不得不大打价格战，而企业依靠价格战并非长久之计。这五家公司均由浙江杭州桐庐县人创立，且创始人都来自该县钟山乡几个相邻的村庄，有人戏称“桐庐一个地方120万人，80万在干快递”。这些同根同源的快递公司，在占据市场垄断地位后，面对依然持续增长的用户需求和加速上涨的成本，共同上调价格也就不

难理解了。

竞争导向定价法

竞争导向定价法适用于价格敏感度高的商品，以市场上相互竞争的同类商品价格为定价基本依据，根据竞争状况的变化确定和调整价格水平，与竞争商品价格保持一定的比例，而不过多考虑成本及市场需求因素。

世界上最大的工程机械设备生产商、成立于1925年的美国卡特比勒公司，对于牵引机的定价方法就十分奇特。一般牵引机的价格均在20 000美元左右，然而该公司却报价24 000美元，每台比同类产品高4000美元，即20%，但它的销路却很好，为什么呢？原来它们有一套说服人的账单：20 000美元是与竞争者同一型号的机器价格，5000美元是产品更耐用而必须多付的价格，3000美元是保修期更长多付的价格，合计价格是28 000美元，如果把折扣定为4000美元，这样24 000美元就是最后的价格。这样一算，加深了客户对该公司产品性能价格比的理解，还使得众多消费者自己宁愿多付4000美元，结果卡特比勒公司的牵引机在市场上十分畅销。在这个报价过程中，卡特比勒公司就充分感知了消费者对商品价值的感受和理解程度，并且参考竞争对手的价格合理定位。

竞争导向定价法还可分为以下几类。

(1) 通行价格：使零售店商品的价格与竞争者商品的平均价格保持一致。

(2) 主动竞争：不追随竞争者的价格，而是根据零售店商品的实际情况以及与竞争对手的商品差异情况来确定。

(3) 密封投标：主要用于投标交易，投标价格是零售店根据对竞争者的报价估计确定的，而不是按零售店自己的成本费用或市场需求来制定的。当商场位置好、产品稀缺、优先上市、品牌形象和信誉度高于对手时，可以考虑定价高于竞争对手，虽然制定更低的价格容易吸引客流，但可能会导致恶意

竞争。

现实生活中，商品定价要受到各种因素的影响，除了传统的成本导向、需求导向、竞争导向以外，在具体的情景中，还会有创新的组合模式，例如：

- 捆绑销售：商品组合定价，竞争对手很难模仿；
- 支付方式定价：通过支付方式不同获得不同的价格或折扣，例如银行卡与现金，通常支付现金会更便宜，因为银行卡需要付给银行手续费；
- 短期特价法：利用价格弹性大的商品短期降价，价格在高-低-高波动，从而刺激消费者的消费欲望；
- 数字游戏定价法：有尾数法（例如9.99元、1999元等）、奇/偶数定价法（这种方法更多是为了造成视觉冲击，例如777、888、999）等；
- 拍卖定价法：主要用于古玩交易、土地交易等大宗商品交易中。

直到今天，企业在定价过程中依然需要不断地结合传统的定价方法与策略，进行策略组合，这些方法并不新鲜。然而，用传统的价格调节手段，其调节过程通常都会花费较长时间。随着数据经济的发展和计算能力的变革，商品的供给和需求相比以往发生着更加剧烈的变化，价格不再仅仅由商品属性或行业特性决定，而是与时间和空间需求息息相关，这时，传统定价策略已经无法满足赛博新经济中的需求的剧烈变化。那么动态定价算法该如何解决这些问题呢？

动态定价的算法基础

赛博新经济带来的最大变化是大数据和机器学习算法。没有在线数据的时候，价格的变化依赖假设模型，大部分时候是拍脑袋决定的，多次试错后才能找到合适的价格，每次试验的真实数据需要上一次试验得出的结果，反

馈机制比较长、收敛慢、试错成本高。有了在线数据以后，动态定价算法可以实时测验。同时，计算能力的增强，可以辅助算法更精准高效地挖掘合理的价格，把算法迁移到不同的应用场景中，通过动态调节商品价格促使其回归价值，进而推动各行各业的市场秩序趋于理性。所以，动态定价作为一种强有力的价格调节机制，其策略空间非常大。

以赛博经济中的平台型企业滴滴公司出行为例，我们可以归纳出一般的动态定价机制。如图4.4所示，滴滴出行周期性地收集实时乘车数据，集合自身运力进行预测和匹配，在运营过程中不断反馈与迭代，进而再次发现价格。

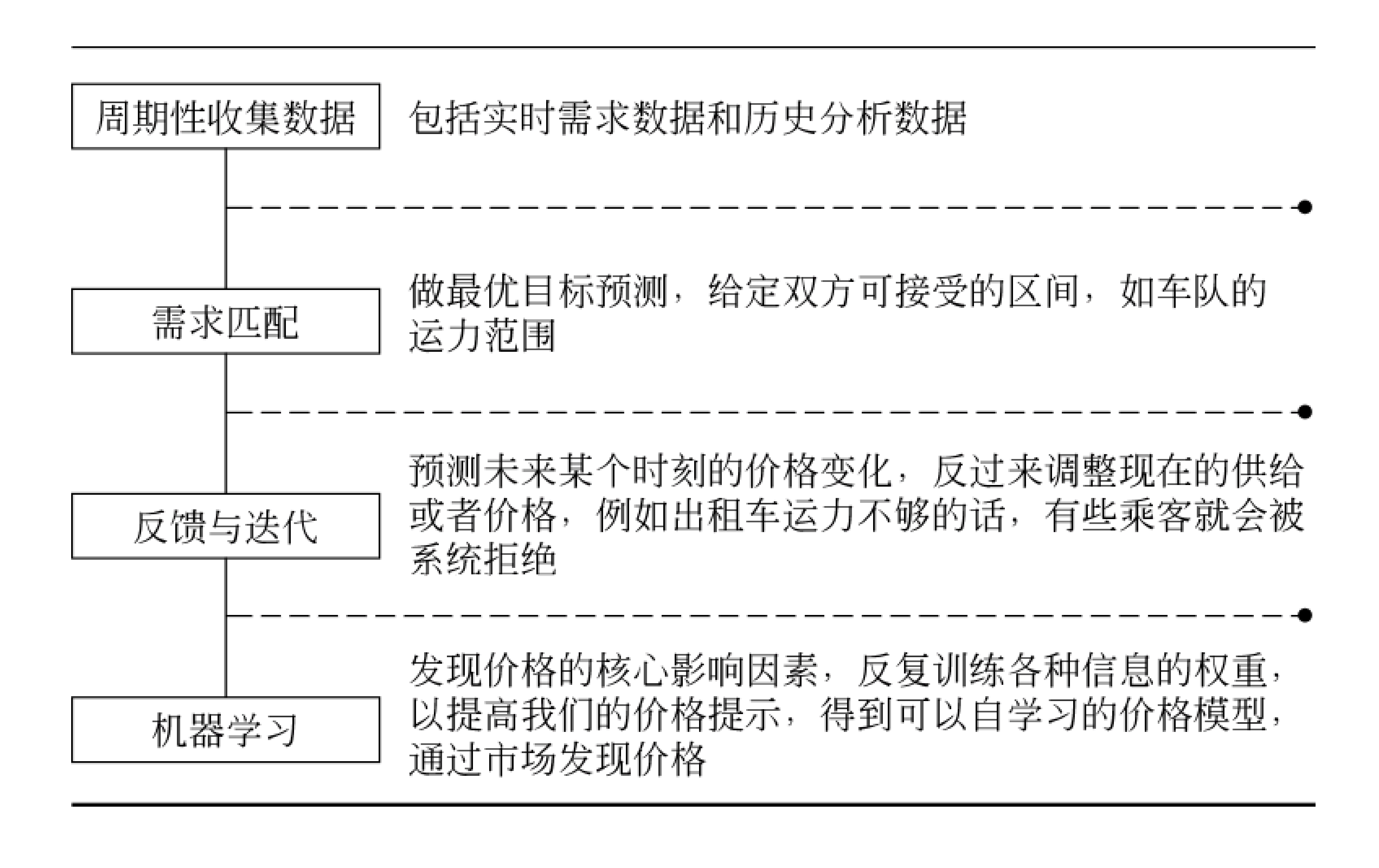


图4.4　动态定价机制流程图

人工智能算法的进步使得动态定价的算法也越来越智能，借助机器学习算法，动态定价算法甚至也可以自我学习和实时修正。

目前，人工智能(AI)已成为全球IT巨头最新的角斗场。谷歌公司旗下的AI科技公司DeepMind去年推出AlphaGo，相继战胜人类顶尖棋手李世石和柯洁，不禁让我们联想起几年前IBM公司旗下的云端AI计算机沃森赢得了美国电视智力竞猜节目《危险边缘》特别版的冠军，击败美国竞猜节目中

最聪明的人脑。DeepMind 和沃森与人类的较量，把人工智能的应用推向各个领域，开启了对大数据分析背后的知识与洞察的争夺战，也将“深度学习”变得广为人知。而深度学习需要的两样东西，正是海量的计量处理能力和海量的可学习数据。沃森背后依赖的是被 IBM 称之为“认知计算”的基本原理，其核心也正是机器学习算法。

下面让我们先看一下动态定价中供需双方的交易点最后应该处于什么位置。

动态调整价格的理论预期值——纳什讨价还价解

设计一个动态算法，首先还是要从作为讨价还价的买卖双方的博弈开始。我们与小商贩讨价还价，是一种最为原始的动态定价行为，它的历史可以追溯到原始的商品交换时期。在没有经济统计数据、技术不发达的年代，商品交换以一种原始的方式进行着，那就是在街头巷陌的市场中，商贩与消费者讨价还价。这个过程是分两步完成的：首先是双方沟通议价，通过交谈的方式将买方和卖方对于此笔交易的心理预期和市场定价等一系列交易信息进行交流，并且通过不断地调整定价策略来达到符合自己意愿的交易价格。最后是成交，双方根据之前的协商定价完成交易。

讨价还价涉及一个大众心理问题，价格需进行讨价还价的博弈互动才能达成，这种博弈，曾因纳什的贡献而被称为“纳什讨价还价”，双方通过讨价还价达成的一致价格，称为纳什讨价还价解。

通常，在参与交易的双方中，卖方有一个心理预期最低值，买方有一个心理预期最高值，买方的最高值应该高于卖方的最低值，否则交易无法达成，而在买方的最高值和卖方的最低值之间的任何一点上都可能实现成交。在赛博新经济时代，我们可以通过人工智能和大数据了解供给能力，同时预测需

求情况，这样在动态定价的设计上，就可以让双方更加快速地达到纳什讨价还价解这一交易点。下面我们对以纳什命名的这一博弈过程做简单介绍，了解双方达到交易的价格点在哪里。

假设有两个人分一大碗冰淇淋，甲先出价（这里的出价实际上是指冰激凌的分配方案，也就是甲和乙分别占多少份额），如果乙接受，冰淇淋按甲的方案进行分配。如果乙拒绝，他将还价，甲可以接受或拒绝。如果甲接受，博弈结束，冰激凌按乙的方案分配。如果甲拒绝，他再出价。如此来回出价，直到一方的出价被另一方接受。这些你来我往的过程可以看作是一个无限次完美信息博弈，这里的完美信息指的是双方都没有任何隐藏信息。第一个参与人在时间点 1、3、5、……出价，第二个参与人在时间点 2、4、6、……出价。但是请注意，天气很热，伴随着每轮议价，冰激凌会不断融化，从而价值越来越小，反映在价格上也会有所降低。也就是说，价值在不断缩水，在经济学中，用贴现因子来表示这种价值缩水的过程，贴现因子通常是一个小于 1 的数，表示价值在不断减少。为了准确计算出最终交易时冰激凌的价值在前面每一轮议价时值多少钱，可以假设甲、乙的贴现因子分别为 δ_1 和 δ_2。

假如，我们用 X 表示甲的份额，$1-X$ 为乙的份额，那么，X_i 和 $1-X_i$ 分别是时间点 i 时甲乙各自所得的份额。这里需要知道甲、乙未来所得的份额折算到当下的价值，前面设定了贴现因子分别是 δ_1 和 δ_2，那么，当博弈在时间点 t 结束时，甲的支付贴现值为 $W_1=\delta_1^{t-1}X_t$，乙的支付贴现值为 $W_2=\delta_2^{t-1}X_t$。两者在无限次讨价还价之后，双方终于就某一价格达成一致意见，即最终得到的均衡解为：

$$X'=\frac{1-\delta_2}{1-\delta_1\delta_2}$$

这就是纳什讨价还价解，其意义在于当双方都知道纳什讨价还价解

存在时，实际上甲方第一次出价时就会按照均衡解直接给出建议，乙方接受，博弈过程结束，并不需要真的重复无限次，这样至少能保证冰激凌大体完整。

在经济学中，讨价还价模型可以应用到很多场景中，最经典的就是商业谈判。原本的数学模型是以冰激凌作为整体进行考虑，现在可以把企业甲和企业乙谈判过程中，出的最低价 a 与出的最高价 b 组成一个区间，而议价过程中的每一个价格都会落到$[a, b]$中来，事实上，谈判的双方就是在这一区间上出价议价的。经过谈判，双方会在价格 c 处成交，而 c 一定处在 a 与 b 之间。由此，我们可以得到新的纳什讨价还价解。

$$X' = \frac{(1-\delta_2)(b-a)}{1-\delta_1\delta_2} + a$$

下面可以看一下这个算法如何在市场中应用。假设公司甲为了改善培训课程的排课体系，并且培养学生们的自主规划能力，拟收购公司乙，但是对于初创企业，市场并没有一个合适的公允价格，于是两家公司进行并购谈判。经过资产评估，课程格子的净资产为 100 万元，根据当时市场状况及商誉等情况，公司乙 CEO 决定出价 130 万元，但是，公司甲认为其价值只有 110 万元，于是还价 110 万元。这里公司乙先出价，公司甲后出价。假定双方的贴现因子相同，均为 0.9，根据模型，可以计算出双方谈判的均衡结果 X' 为 120.53 万元，它们最终会以这个价格交易。

然而，这是根据公式计算出来的理想均衡结果，双方的成交价格还存在许多客观或主观因素，不一定等于 X'，例如，如果公司丙此时也对公司乙有兴趣，于是抬高市场价格，或者公司甲内部有一款类似软件，可取代公司乙的部分功能，并不需要全资收购，这些都会影响最后的成交结果。但是，这个经典模型无疑对未来的成交价格具有重要参考作用。

动态定价在新经济领域的应用

策略性思维是在不断弄清对手的过程中，战胜对手的一门艺术。

——埃维纳什·迪克西，《策略性思考》

机票价格到底是如何制定的

航空公司飞行线路图错综复杂，我们日常乘坐飞机的交通成本往往是旅行费中的“大头”，但人们常常发现机票价格变化随机得令人摸不着头脑，有时高得离谱，有时突然大幅打折，长途机票经常比短途机票还便宜。CNN 有过一次相关报道，这些看似无序的现象背后，其实是有规律的。为了实现收入最大化的目标，航空公司往往实时动态地调整票价，这种方法称为航空公司收益管理。它们不仅会参考成本、供给与需求等方面，还会利用先进的软件，将公司的全球航线布局、乘客个人偏好等因素考虑进来，以实现科学定价。

航空公司定制个性化价格

历史上的大多数时期，航空公司处于严格监管之下，国际航线通常由相关国家的国有航空公司运营，几乎不存在市场竞争，买机票往往需要一笔巨款。打折机票并非不存在，但通常有很多严苛的附加条件。1978 年，在全球贸易自由化趋势下，美国出台法案放松对航空业的管制，于是，产业结构和票价也发生了很大变化。网络的发展与计算机成本下降，使收益管理的先进性

达到了全新水平，航空公司的收益管理也变得竞争激烈而日益复杂。

人们都知道经济舱、商务舱和头等舱的票价不同，但其实，航空公司把机上的座位细分成了几十个部分，并根据机票售出情况随时调整每个类别的座位数量。CNN 称，航空公司的终极目标是了解客户，并提供完全个性化的定价服务，“例如，伦敦到马略卡岛的航线明显是度假休闲路线，航空公司假设旅客会提前几个月订票，那么这条航线的机票一开始会相对定价较高，然后根据市场反应进行调整。而对于伦敦到法兰克福这样的典型商务航线，航空公司一开始可能以低价确保机票能够售出，然后在起飞前最后的时段里大幅提高票价，将机票卖给刚性需求的客户。”此外，航空公司收益管理系统不仅考虑机票本身，也越来越多地考虑旅客可能带来的辅助收入，这将成为日益增长的利润来源。

为了与廉价航空公司争夺客源，美国联合航空公司曾推出比经济舱更便宜的特价机票，购买这种机票的乘客可以享受到与普通经济舱乘客同样的食品、饮料、娱乐和无线网络服务，但在航班起飞当天才能分配座位、最后登机，且只能携带背包、电脑包等能放在座位下方的个人物品。《纽约时报》称，有些乘客只想安全到达目的地，并愿意为更优惠的价格放弃一些服务，航空公司则允许乘客在购买廉价机票的基础上多花一些费用，获得额外的服务，如行李托运、更宽敞的座位等。如果航班在起飞前仍有空位，将其低价售出是不小的诱惑，但航空公司需要考虑机票价格与品牌和长期业务之间的微妙平衡。如果起飞前总有低价票放出，就会严重损害公司品牌，导致高附加值乘客流失，商务舱乘客也可能不愿再购买全价票，因为他们知道有机会在最后关头以很低的价格升舱。为此，以色列的 Bidflyer、美国的 Plusgrade、澳大利亚的 SeatFrog 等创业公司纷纷推出航空竞价软件，航空公司可通过拍卖将升舱服务卖给出价最高的乘客，并以此了解乘客愿意为额外服务支付多少钱。

买机票将成为机器人大战

虽然机票定价有规律可循，但对许多乘客来说，价格波动得如此随机仍令人感到焦虑，他们担心买不到价格最优惠的机票，甚至被“宰”。航空公司有一系列策略来增加收入，旅客也有自己的对策。据CNN报道，美国优惠机票预订应用Skyscanner、Kayak等，可以帮助旅客搜索超过1000家航空公司的上百万条航线，并在几秒钟内找到价格更低的航班，同时为旅客实时监控特定航班的票价情况并发出提醒。据“美国之音”电台网站报道，与以往的机票垂直搜索和比价网站不同，凭借独有的票价预测技术，美国加州FLYR公司不仅承诺帮助旅客在最佳时刻预订价格最低的机票，还与TripAdvisor合作推出了“锁定票价”服务。用户在该网站上选择航班后，FLYR将在7天内确保他们预订的票价不受价格增长的影响，如果用户在7天内搜索到了更合适的票价，变更选择也不需要额外缴纳费用。

“根据大量的数据资源，我们可以准确地预测机票价格波动的可能性，为旅客提出最佳出行方案。如果我们认为票价将在不久的将来下跌，会建议乘客等待一段时间。”FLYR创始人、荷兰企业家Alexander Mans告诉CNN，在航班出发前30天，机票价格有60%～70%的概率下调。Hopper是另一家专业的机票价格预测公司，它使用大数据提前12个月预测票价，其手机应用程序已被下载超过800万次。“我们的系统每天监控60亿到80亿张机票，我们的数据库有过去5年的历史票价。”Hopper创始人兼CEO Frederic Lalonde自豪地称，他们的算法能在航班出发6个月前准确预测机票价格，误差不超过5美元，准确度高达95%。大数据和人工智能技术的发展，有望成为机票定价领域的主流，未来购买机票将是一场“机器人大战”。这并不是坏消息，它可能为旅客带来更好的选择和更有效的预订流程。

Uber 的动态定价策略

在 2016 年 CES 国际消费类电子产品展览会期间，Uber、Lyft 等打车应用也悄然在美国拉斯维加斯所在的内华达州下调了其服务价格，助力 CES。在 CES 举办之前的 2015 年 12 月 8 日，Uber 就已经将其在内华达州的打车价格下调，并且不再上涨，整体价格已低于当地的出租车行情。随后，Uber 的竞争对手 Lyft 也效仿其降价方法，并且 Lyft 发言人赛利亚・布雷森还表示，在 CES 期间不会上调高峰期 200%的涨价上限。与此同时，Uber 还联合车载设备厂商 Vinli 以及无线运营商 T-Mobile，在 CES 展会期间面向当地乘客提供 Wi-Fi 服务。虽然这些举措是 Uber 在 CES 着力推介自己的行动，却为 CES 展期间的交通调度带来了便利。

现在已经被滴滴出行收购的 Uber 中国，在最开始进入中国市场时，便加入了出行补贴的战局，而与国内打车公司不同，动态定价策略一直是其核心策略之一，虽然那些对价格敏感的消费者对此并不满意，但这不但不会影响 Uber 的销售，反而这种符合实时供求关系的溢价算法提高了 Uber 的整体收入。正如他们自己所说“溢价不是计划好的，是依据供求动态平衡”。

Uber 的峰时价格调整策略

2012 年初，Uber 位于波士顿的研究组发现，每到周五和周六凌晨 1 点左右，会出现大量的“未满足需求”。导致这种现象的原因是在这个时段，大部分司机退出 Uber 系统，准备收工回家，而恰恰这会很多人刚准备回家，造成了瞬间的供需不平衡，在最需要用车的时候却叫不到车，用户的抱怨与日俱增。于是 Uber 设计了一个方案，在高峰期（午夜到凌晨 3 点）适当提高每次

乘坐的单价，看是否能吸引更多司机。仅仅两周后，他们就得到了非常好的反馈，在该时段的提价，使得服务车辆的供应量增加了 70%～80%，几乎满足了三分之二的“未满足需求”，这绝对是个重大突破。看来在出行领域，由于供应量的弹性非常大，在市场价格调高后，司机确实更有动力守候在午夜时分。

这个调查成功开启了 Uber 的动态定价策略，随后动态定价便正式应用在所有高峰时段。Uber 动态定价的算法也十分智能，当用户等待时间的上升趋势比较陡峭时，便会触发该算法。其实，这里采用的动态定价核心想法很简单，要解决供求不平衡，要么增加供给，要么减少需求。动态定价则成功地从这两个方面影响了供求关系。

供求曲线模型是经济学最基础和最核心的模型。要分析，就要根据 Uber 的业务模式来确定其供求。图 4.5 给出了供给与需求的关联模型。

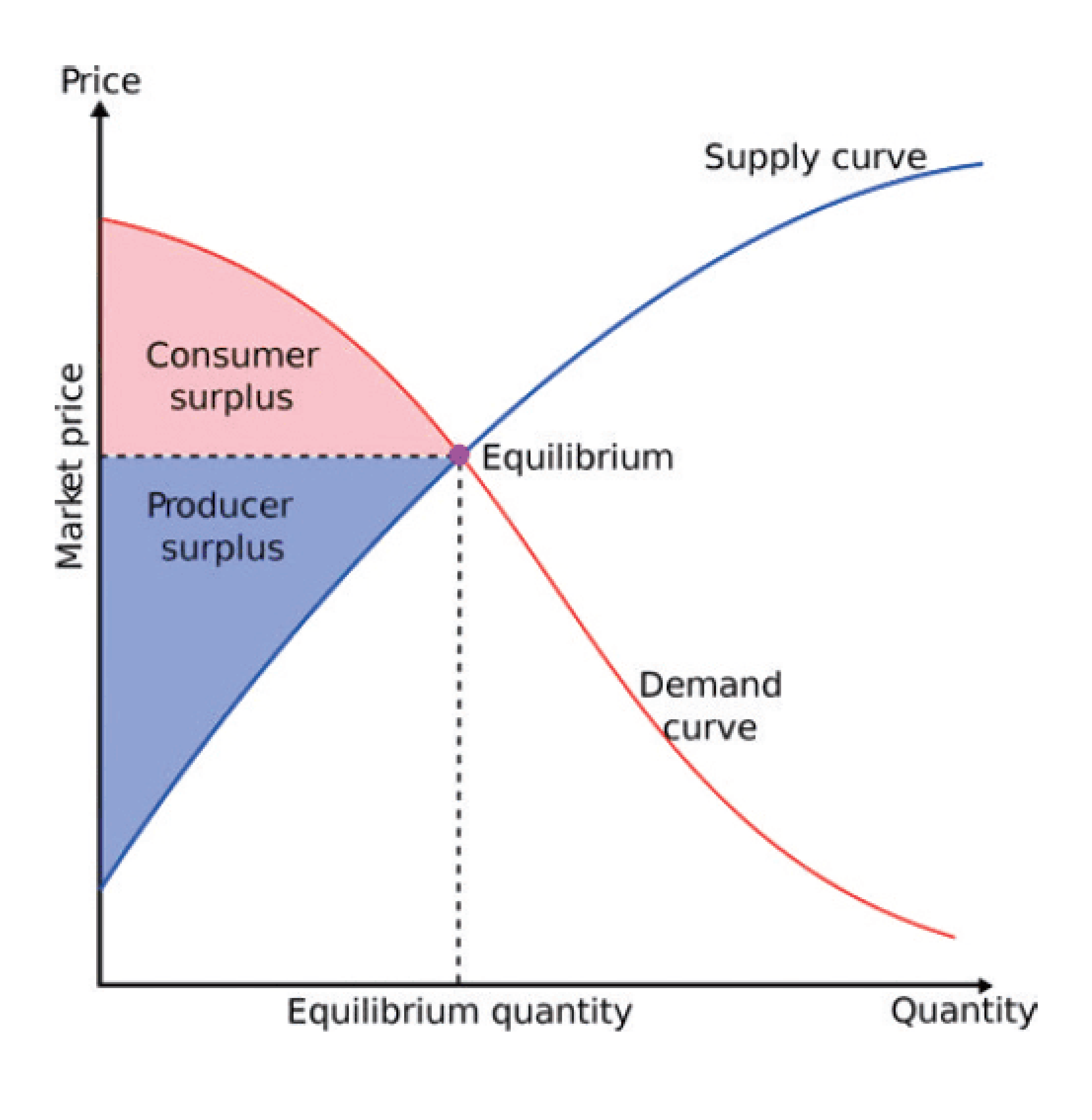


图 45　供给与需求模型

首先，如上文中的波士顿实验，证明出行市场的需求和供给都是高弹性的。

就需求方而言，在两个方向上都具备高度弹性。其一，当价格升高后，直接使需求量相应减少。其二，当价格降低后，需求量也会立即增加。

市场化在 Uber 的业务上体现得非常灵敏，“看不见的手”将资源进行最优化的配置。这种现象并不难解释，因为在 Uber 这个市场里的参与者，都是独立个体，可以认为是“理性人”，所以他们的行为可以准确地被市场规律描述。

Uber 动态定价模型中的供求关系也是非常直观的。

当需求大于供给，算法会自动提高价格，减少需求提高供给，使得供需达到一个动态平衡。这个过程持续不了多久，因为当供给逐渐大于需求时，价格又会恢复到初始水平。这个过程循环往复，始终维持着平衡。试想如果需求增加，而不升高价格，会发生什么？用户等了好久都没叫到车，未满足需求井喷，用户不满意，将卸载其软件，再也不用。如果采用动态定价，从图 4.6 来看，如果需求增加（图中从 D_1 到 D_2）或者供给减少（图中从 S_1 到 S_2），当然也可能同时发生，这时，新的平衡点 Q_2 和 Q_1 比，价格是上升的。

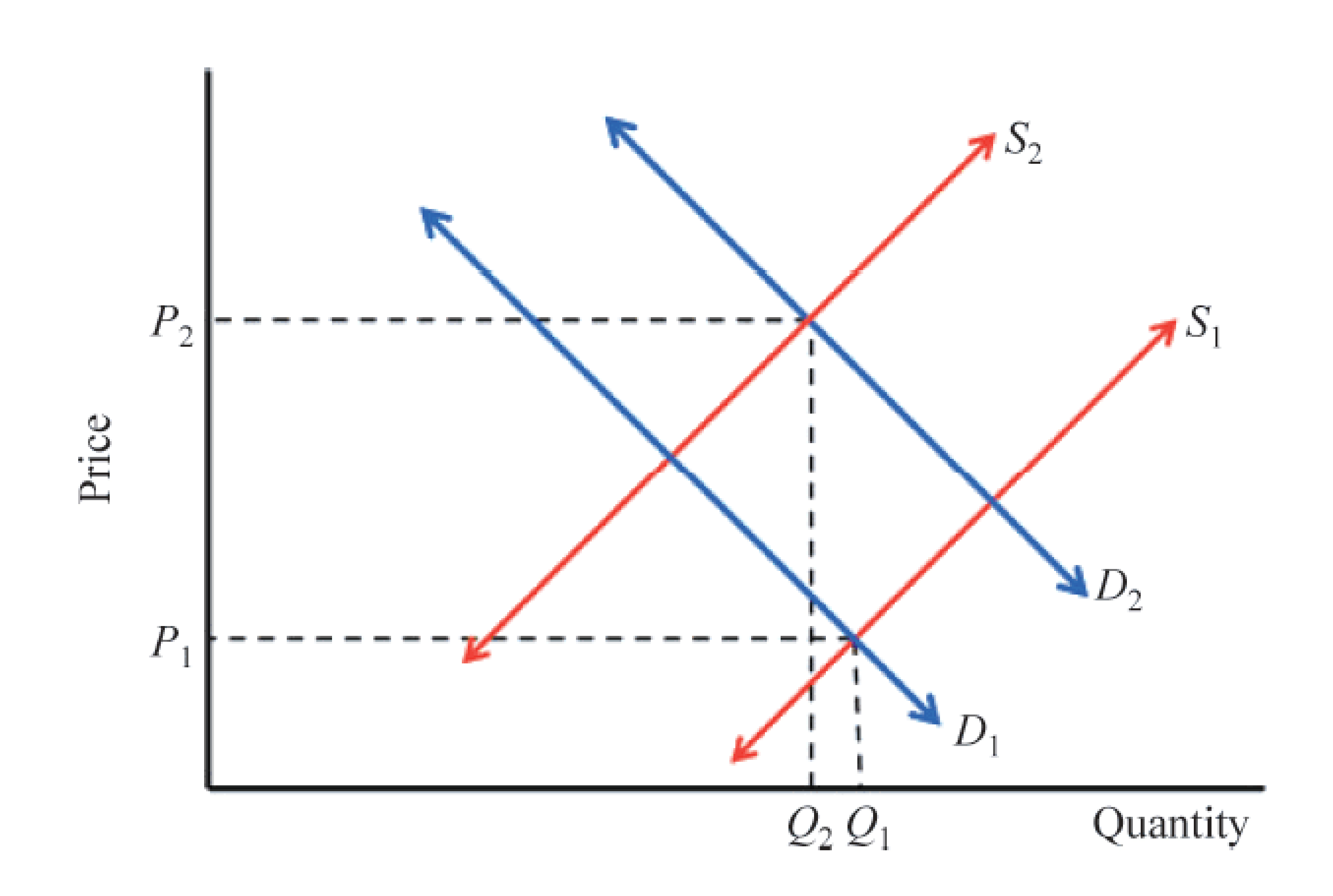


图 4.6　动态定价的供需分析

Uber 与酒店、机票、租车

很多行业都在比较成熟地使用动态定价，例如酒店、航空公司、租车行，高峰期也与 Uber 类似，尤其是节假日。酒店在新年夜的价格往往比平时或周末都要高出一两倍，在无法提高供给的时候，提高价格也是能被大众所接受的举措。

唯一不大一样的是，像酒店、机票，它们的供给是固定的，无法提高，而 Uber 不同。对酒店来说，供给是刚性的，因为无法临时造所房子出来，而 Uber 的司机供给弹性则大得多，司机可以收工回家，也可以继续服务。

大家知道，绝大多数运用动态定价的企业是有库存的。它们的特点是“有限、随时间消逝”，即一趟航班的座位数、一个酒店的房间数、一个球场或剧院的座位数都是固定的，一旦飞机起飞、比赛开始、演出拉开帷幕，时间过去了却还有空座或空房，那么这部分收入也就永远失去了，因此它们要尽可能快地售清库存，并尽可能多地从需求中获取利润。可以说，传统的动态定价都是通过价格的变动来调整需求，目标是适应供给，或许称为“基于时间因素的差异化定价策略”更合适。

Uber 则不同。Uber 不拥有任何一辆车，也无法强制任何一个司机服务，但任何一个有车的司机都有可能安装 Uber 司机端软件，换句话说，Uber 的“库存”是可多可少的。人们搭车需求最强的时刻，对司机来说往往也是驾车体验不甚愉悦、甚至危险系数颇高的时刻，例如早晚高峰、有暴雨或台风的夜晚。在这些情况下，若无激励机制，上线服务的司机数即供应量自然就会减少。因此，使 Uber 定价区别于其他行业做法的特殊之处，不在于它限制了需求，而在于它调动了供给。以前想打车也付得起车费却无车可打的人，现在可以顺利叫到车了。最终的结果是让更多人的需求得到了满足。

对 Uber 来说，在需求大量增加时，供给曲线左移，需求曲线右移，这时

需要价格作为催化剂来达成二者的平衡。影响供给的还有个因素，就是可替代性如何。新年夜的大单千载难逢，有的用户会预定一辆车来独自享受，价格甚至可能超过 1000 美金一晚，这种情况下，有些在家里无事可做的司机也是会很乐意出来接一单的。

所以，Uber 的价格，实际上是基于“供需关系”这个最基本的逻辑，依靠动态定价算法实时计算得来的。例如在北京五道口的某个时刻，当前有 100 个司机和 100 个乘客，供需平衡，那么采用“基准价”就可以了。过了半小时，100 个司机还在，只有 80 个乘客了，这时，Uber 动态定价算法就开始下调“基准价”，把那些本来准备坐地铁或骑共享单车的人吸引过来。同时，也顺便减少一点供给，有些司机认为这个时间段太便宜，干脆去歇一会算了。两个小时以后，晚高峰来了，100 个司机没变，有 150 个乘客了，算法开始自动加价，1.5 倍、2.3 倍、3.8 倍不断上涨。觉得贵的乘客自然会去骑单车，价格的上涨顺便还能把周边等活的 Uber 司机也吸引过来从而增加供给。

最具挑战环节：附近无车可用

关于 Uber 的定价，媒体舆论也吵得纷纷扬扬，使得 Uber 不得不重新考虑其加价条款。越来越多的用户抱怨在很多地方都无车可用，丝毫不实用，也无可靠性。最差的一种体验就是刚打开 Uber，就提示无车可用，如图 4.7 所示。有人说，至少应该在没车可用时保持平价，好站在用户这一方，抚慰这些没打到车的人。其实事实不是这样，在高峰期，任何一种交通工具都是超负荷运转的。地铁、公交都是这样，都无法提供可靠的服务。这时 Uber 更倾向于让更多的用户能够叫到车。与其让用户无车可用，不如让部分用户对定价持有异议似乎更容易接受。

不理解 Uber 动态定价的用户，其实是没有理解 Uber 作为一个市场平台的本质。主流的平台都会用供需模型来调整供应量。这也是 eBay 拍卖最

初的做法。StubHub、Airbnb、Homeaway 也都是这么做的。Google Adwords 的定价算法也是以此为基础。正是动态定价在市场上如此广泛的应用，给了 Uber 信心。最大化用户的利益，最优化市场资源的配置，只能通过动态定价来实现，即使有时部分用户会很不理解。在车少导致打不到车的情况下，大家在心理上更希望把打到车的人归结为“运气好”，而不是“他有钱”，而且在极端天气到来的时候提高价格，会让人们本来把对某个司机发灾难财的抱怨转嫁到整个公司的头上。

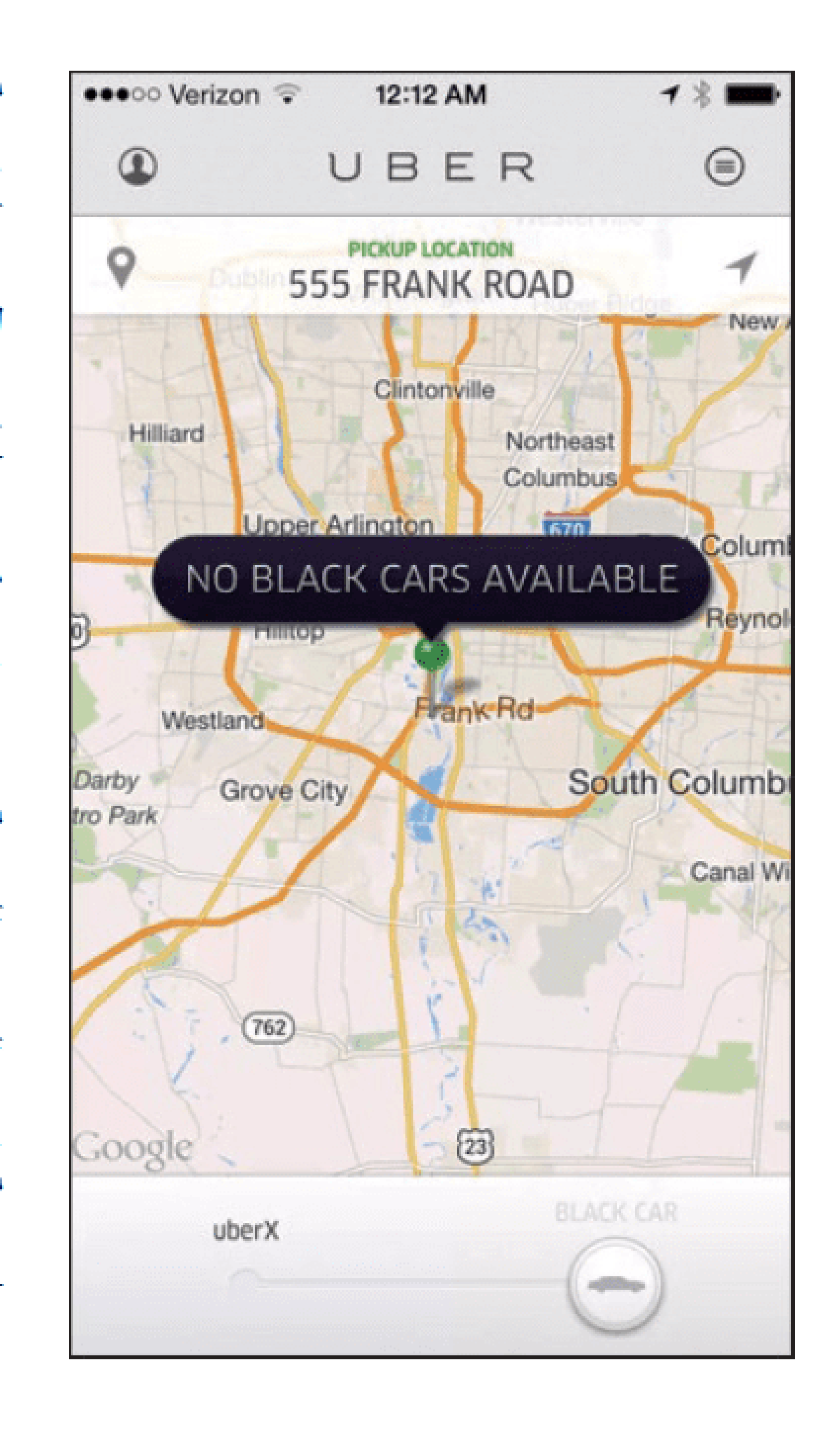


图 4.7　Uber 客户端显示界面

Uber 对经济学原理的应用无可厚非，供需原理也的确能够发挥杠杆作用，但是归根到底，在算法还无法达到真正的智能之前，比如即时判断车变少的原因，完全依赖于大数据和算法总是会让人不太愉悦。

Uber 的定价算法基于大量、即时的数据进行建模和优化，引入了“时间-空间”双维度。这种实时性，赋予了司机估算需求的能力，这在传统出租车行业是无法实现的，Uber 还能在系统给出的建议指导下，识别出回报率最高的接客时间和地点。此外，不同城市的价格弹性各不相同。为实现对每个城市的“量身定制”，Uber 建立了大规模的计量经济模型和数据库，量化不同城市中乘客与司机对价格的敏感度、候车时间等相关变量，并随变量的改变即时调整算法，以适应不断变动的市场情况。

另外，Uber 的动态定价严格意义上来说应该是“基于需求的动态定价”，因为在同一时间里，并不是所有产品都面临“上下班高峰”或者“暴风雨”的外

部条件，而所谓“传统的基于时间动态定价”的产品，例如迪士尼乐园门票，同样是销量导向的产品，同一时间所有产品都面临同样外部条件。Uber 的“共享经济”模式与传统的商业模式有本质的区别。

2014 年 12 月，澳大利亚悉尼的公共场所发生了一次匪徒持枪劫持事件。路人都四散奔逃，没车的人只能叫 Uber 服务。由于需求陡增，Uber 动态定价机制自动决定提价。当地的 Uber 价格瞬间涨到原有价格的四倍。事件过后，澳大利亚的社会评论家撰文抨击 Uber 跟恐怖分子一起干起了“趁火打劫”的买卖。各大社交网站也是一边倒地怒斥 Uber 的“无良举动”，在重压之下，Uber 同意提供最多 200 澳元的退款。虽然做出了退款举动，但 Uber 发言人还是对涨价之举做出了解释：打车服务的车费是根据算法计算得出的，该算法会对需求自动做出回应。

实际上，Uber 可以看作一个双边共享平台，让想坐车的乘客在合适的时机打到车；也可以看作是双向资源匹配问题，价格则是解决匹配问题的杠杆，基于差异化的时间和需求，动态调整行程和价格，可以有效调度双边资源，当实时数据积累到一定程度，再利用大数据技术更精准地预测路况，从而智能调整价格。在出行这样一个长期以来定价几乎完全受人为控制的行业，Uber 通过动态定价，让市场复位，重新发挥“无形之手”的力量，这是它最有价值的地方。有人曾经发表长篇文章力挺 Uber 应该获得诺贝尔经济学奖，认为 Uber 有史以来第一次解决了人类经济学上最大难题之一，即定价问题。这个说法也许有点偏激，但是动态定价如此深入人心并让人又爱又恨的确是 Uber 的功劳。

Airbnb 的动态定价算法

Airbnb 作为世界上最大的居住信息共享服务型平台，有大量的外出旅

游人士和家有空房出租的房主，不同于电商平台，它只是告诉用户哪些商品已售出或评价如何，而无须关心买家卖家在哪里。Airbnb 中数百万房源都是独一无二的，它们有自己的地址、大小和装饰，而顾客在接待、饮食或旅游引导方面的要求也不尽相同，再加上天气的季节性变化等，都会让定价问题变得更加复杂。

静态的定价机制

Airbnb 的产品主管 Dan Hill 曾设计了这样一套定价机制：当一个新房主开始在网站上添加一个房源的时候，系统提取房源的关键属性，查看在这区域中有相同或相似属性的且被成功预订的房源，同时考虑到需求要素和季节性特征，提供一个居中的价格提示。Airbnb 选择了三大类型的数据来设置价格：相似性、新旧程度和位置。图 4.8 给出了一个例子，说明季节性需求和当地的活动会导致房屋出租价格起伏，在美国得克萨斯州奥斯汀市，在 South by Southwest Conference & Festivals[①] 和 Austin City Limits Festivals 音乐节期间，房屋出租价格会上涨。在定价算法的早期版本中，算法以房源为中心绘制一个不断扩大的圆圈，考虑在房源位置附近不同半径上与其特征相似的房源。

随着时间的推进，Dan 也在不断改进他们的算法，目前已经能够考虑数千种不同的因素，并在非常精细的水平上区分地理位置。但该机制仍然存在两个不足。其一，它给出的这些价格提示是静态的，事实上，在了解了当地的一些活动和旅游的季节性变化之后，它应该在一年的不同季节中为相同属性

① South by Southwest Conference & Festivals：西南偏南大会与艺术节，简称为 SXSW，是每年在美国德克萨斯州首府奥斯汀举行的一系列电影、交互式多媒体和音乐的艺术节与大会，目前已经成为世界上最大规模的“音乐＋电影＋科技”盛会。

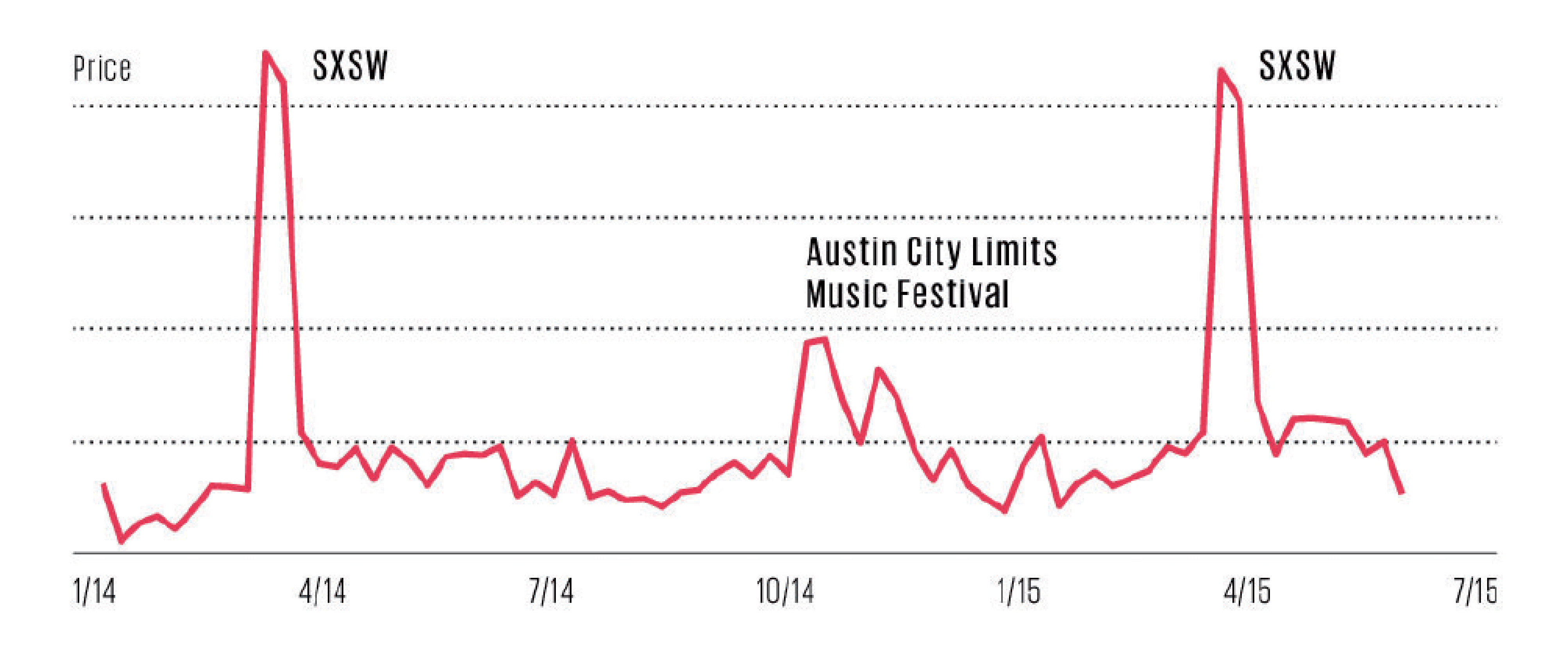


图 48　房价的季节性变动

的房源建议不同的价格。但它并没有这样做,作为对比,航空公司则会在日期临近时改变机票价格,订单减少时将价格下调,在市场升温时将价格提高。另外一个不足是,工具本身是静态的。事实上,当工具能够挖掘到前所未有的历史数据时,它的价格提示有所改善,但算法本身并没有变得更好。

机器学习的动态定价

在动态定价方面,我们的目标是希望能够为每个房主,针对他们的房子计划出租的日期,每天给出一个新的定价提示。前面说过,航空公司动态调整机票价格已经几十年了,而且常常是实时的,以试图确保最大的满座率,以及每个座位卖出最高的价格。酒店业也是这样,随着连锁的规模变得越来越大,酒店的业务数据量不断增长,酒店营销也被搬到网上,使得连锁酒店每天可以多次变动价格。因此,Dan 的团队开始对动态定价加大开发力度,随着历史数据的不断积累,充分挖掘它们的意义越来越大。让算法自身不断改进非常困难,尤其是希望系统给出的价格提示具有说得通的理由。在某些情况下,Dan 希望算法能够有自己的“思维过程”,能够从数据中学习而不断提高水平。机器学习几乎是这类系统的必然选择。Dan 选择了一个分类机器学

习模型，使用房源的所有属性以及当前市场的需求，然后预测其是否会被预定。系统计算价格提示基于成百上千的属性，如是否包含早餐、房间是否有一个私人浴室等。通过将价格提示与结果进行比较，对系统进行训练。考虑房源是否以一个特定的价格被预订，将帮助系统调整其价格提示以及评估一个价格被接受的概率。当然，房主可以选择比价格提示更高或者更低的价格，然后系统也会对估计概率做相应的调整。系统之后会跟踪房源在市场上是否被成功预订，并使用这些信息对未来的价格进行调整。

接下来机器学习就开始发挥作用了。通过分析哪些价格提示获得了成功，Airbnb 系统开始调整不同房源属性的权重。一开始会做一些假设，例如地理位置非常重要，而是否有热水浴缸就没有那么重要。他们保留了以前的定价系统中所考虑的某些房源属性，同时又添加了一些新的属性，如“预订日期之前的剩余天数”，这些信息都会对动态定价产生影响。所有新的信息被考虑到模型中，都是通过对历史数据的分析，表明它们与动态定价有相关性。例如，某些照片更可能吸引预订。总的趋势可能会让你大吃一惊，时尚、明亮的客厅的照片，虽然易于得到专业摄影师的偏爱，但相比于用暖色调装饰的、舒适的卧室的照片，它们并没能吸引更多的潜在客人。随着时间的推移，系统能够为每个定价提示产生一个各种因素及相应权重的列表。

Dan 的系统也在不断地调整地图以反映街区边界的变化。因此，系统并不是依赖于当地的地图，例如，一个当地的地图可能告诉我们波特兰开拓者队的恩光街区在哪个地方结束，里士满街区从哪个地方开始，但这并不是预订者所关心的，Dan 依靠一个城市中房源的预订和价格的分布数据来描绘各种曲线。这种做法也让他们发现了以前所没有意识到的“微街区”。这些地区可能有大量的流行的房源，但它们并不一定与标准的街区边界相匹配，或者可能存在一些局部特征，根据这个特征，将一个较大的传统街区分为一个个小的部分可能更加理想。图 4.9 就是根据这样的方法

绘制出的伦敦“微街区”。



图 49 伦敦的“微街区”分布

今天，这些工具为来自于全球的 Airbnb 房源提供价格提示。但是这些工具除了帮助潜在的房主为在线出租服务更好地设置合理的价格之外，事实上还可以做得更多。所以，Dan 的团队将这些工具所基于的机器学习平台(Aerosolve)作为一个开源工具发布。它将为那些还没有接触过机器学习的从业人员提供一个很好的范例。

电商动态定价掀起价格战

动态定价一度作为收益管理的同义词，不仅在航空领域、出行领域长期推行并取得巨大效益，在电子商务领域也早已大量普及。

华尔街日报的最近报告表明电子零售商的定价策略：根据价格研究公司 Decide Inc 的数据，一个通用电器变频微波炉在一天内价格会变动 9 次。在 Amazon 网站上，价格在 744.46～871.49 美元浮动；而百思买的价格在 809.99～899.99 美元浮动，Amazon 价格上涨的时候百思买会跟着上涨；Amazon 价格降低时，百思买也跟着降低。这些例子就是动态定价，且已经被频繁用到 Pricegrabber、Nextag 这样的新型网站以及元老级的 eBay 和

Amazon 本身，目的只有一个：与竞争对手更好地竞争，通过差价促进销售。

感恩节后的“黑色星期五”是美国一年一度的购物狂欢日，为了在低价的同时保障利润，利用软件系统监控对手并且每日多次调整价格已经成为必不可少的手段，如图 4.10 所示。在消费者踏破卖场的门槛，争抢打折商品的时候，传统零售店和电商却在绞尽脑汁展开价格战。对于顾客来说，网上比价后再也不用彻夜排队。所以零售商必须第一时间提供最低的价格。随着电子商务的崛起和各种数据财季分析工具的诞生，零售商对竞争对手的反应前所未有的快和精确，在竞争对手出价后数秒就马上跟进。传统零售商被迫转型，跟随电商的弹性定价。价格战不再只是拼低价，而是成为信息挖掘和策略的系统比拼。

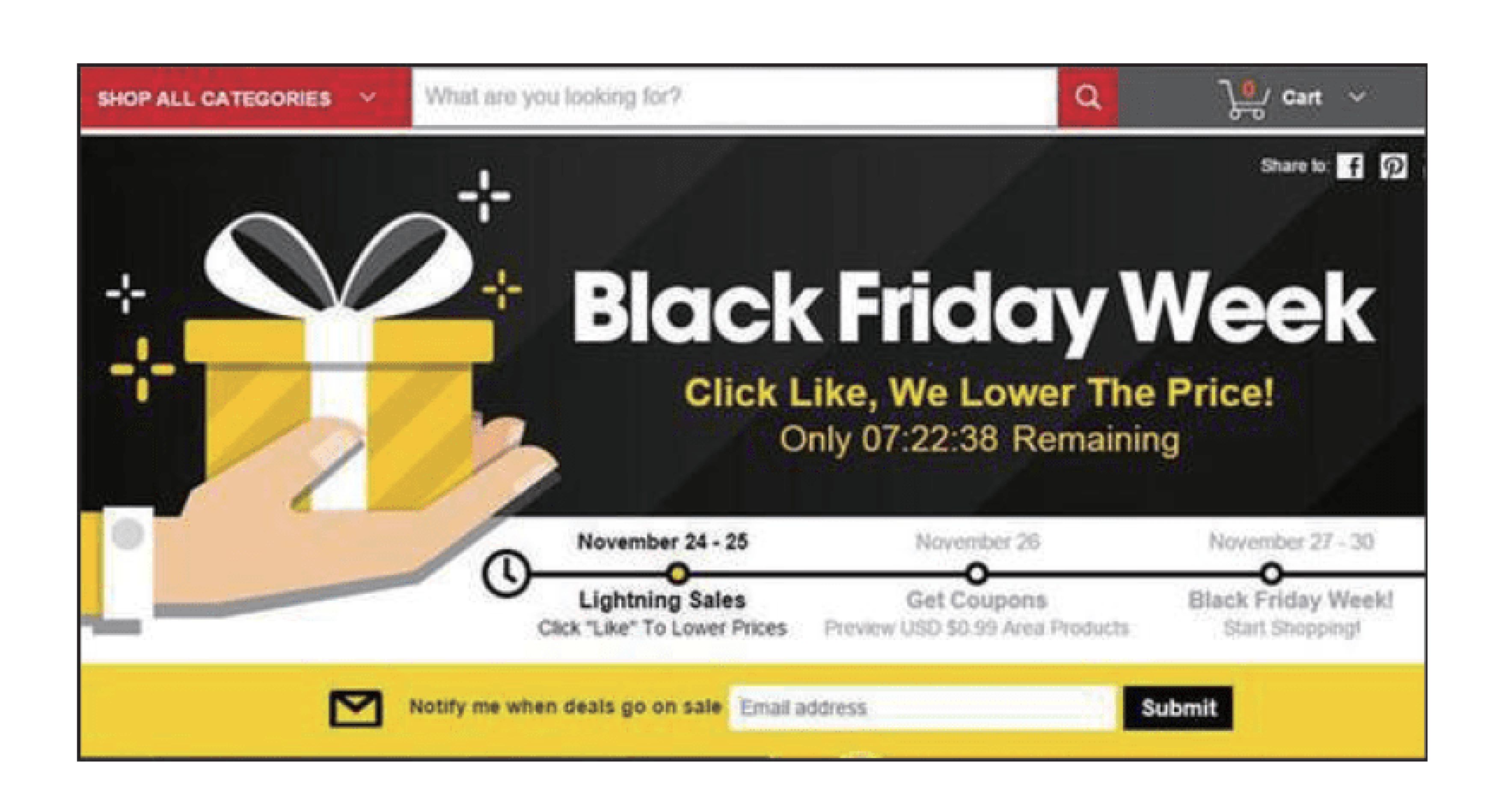


图 4.10　“黑色星期五”电商促销

排名机制对价格战的推动

价格调整最频繁的是在亚马逊上销售产品的网店。亚马逊鼓励零售商之间展开无情的竞争，争夺搜索结果的榜首位置。例如儿童服装店 Cookie's，为了保住在亚马逊排名中的领先位置，卖家使用了软件，每 15 分钟修改一次价格。店主法拉克说，他在亚马逊上销售的服装经常比他在纽约布鲁克林

的商店卖得便宜。“对于新到货的当季新颖款式,我们还没来得及好好营销,就已经在以低于自己预期的价格销售了。”对于那些在亚马逊上销售产品的零售商来说,拥有最低的售价是跻身令人垂涎的“购物推荐榜”的最快捷径。而购物推荐榜或默认商品榜上的产品会有 95%的机会被买家选中。

算法软件加快价格迭代速度

提供价格调整软件的 Mercent 公司称,它们的软件一小时内可以修改 200 万件商品的价格。该软件会根据各种不同因素来定价,例如竞争对手的价格、竞争对手的运输价格以及季节销售额等。零售商自行设置价格调整的时间频率、要跟踪的产品和可以忽视的竞争网站。价格变化最频繁的是家用电子产品、服装、鞋子、珠宝和洗涤剂、剃须刀片之类的居家必需品。Mercent 公司 CEO 埃里克 · 贝斯特(Eric Best)说:“从长期来看,这意味着价格不再是单纯的价格。”

网店 Cookie's 的店主法拉克说,频繁修改价格大大促进了销量,但是也需要注意价格底线。他首先在软件里设定了与竞争对手的价格优势比例,接着他设定了一个不能逾越的价格底线,然后他将竞争对手限定为那些在亚马逊五星评级系统里至少获得两星以上评级的商家。

根据电商比价网站 Decide. com 提供的数据,迄今为止,买家在价格游戏中输赢参半:大约一半的价格修改是降价,一半是提价。Decide. com 对商品价格进行全时追踪,从而确定最佳的购物时机。价格变化有可能非常剧烈。据 Decide. com 网站透露,2017 年 6 月,亚马逊上的零售商在一天之内对一款三星 43 寸等离子电视的价格进行了 4 次修改,价格变动区间为 398～424 美元。中午时分,百思买将这款电视的售价从 400 美元提至 500 美元,之后又回调了价格,而在线电子产品零售商新蛋公司(Newegg)早间将该款电视的价格从 500 美元提高到 600 美元。亚马逊、百思买和新蛋都拒绝对此予以

置评。

收益管理≠动态定价

收益管理是对有限资源和服务充分利用的一种商业模式，其目标是最大化企业获利。由于一定时段内的供给是特定的，如果你让它空闲，对于企业来说就意味着损失，因为企业的固定营业成本并不会因此减少或消失。

然而，最大化收益是动态定价的目标吗？对此，亚马逊曾做过一次差别定价实验。2000 年 9 月，亚马逊选择了 68 种 DVD 碟片进行动态定价试验，根据潜在客户的人口统计资料、在亚马逊的购物历史、上网行为以及上网使用的软件系统确定对这 68 种碟片的报价水平。例如，名为《泰特斯》(*Titus*)的碟片对新顾客的报价为 22.74 美元，而对那些对该碟片表现出兴趣的老顾客的报价则为 26.24 美元。通过这一定价策略，部分顾客付出了比其他顾客更高的价格，亚马逊因此提高了销售的毛利率。但是好景不长，差别定价策略实施不到一个月，成百上千的 DVD 消费者通过互联网知道了此事，导致那些付出高价的顾客怨声载道。除此之外，这次事件曝光后，消费者和媒体开始怀疑亚马逊是否利用其收集的消费者资料作为其价格调整的依据，引起了公众对隐私保护的担忧。为挽回日益凸显的不利影响，亚马逊的首席执行官贝佐斯亲自出马，进行危机公关，答应给所有在价格测试期间购买这 68 部 DVD 的消费者以最大的折扣。至此，亚马逊的“身份定价”试验以完全失败而告终，亚马逊不仅在经济上蒙受了损失，而且声誉也遭受了影响。

因此，在动态定价的设计上，不能纯粹以收益最大化为目标，而是要以尽量更多地考虑供需双方的满意度，让双方都认为公平和值得信赖。否则，一个买廉价用品的客户会以低价得到一个产品，而经常买高价物品的顾客却要为此产品支付更高的价格，这虽然对供方有利，但会令需求方蒙受损失，这样做的结果就是平台的失信，从而最终令每一方都无利可图。

动态定价缓解旅客运输压力

在日本，为了方便游客在东京和大阪之间的交通，铁路部门推出了一款“7 日票”，即 7 日日本全国 JRPASS 周游券，折合人民币 1600 元左右，可以不限次在两个城市之间往返乘坐。图 4.11 给出了旅游部门的推荐线路，从东京到大阪 7 天的行程，如果不使用 JRPASS 的话，需要花费交通费总计 1720 元左右，如果想去别的地方价格会更高。但如果使用 JRPASS 全国券，仅需 1600 元就可以走遍 7 天行程的每个地方，同时还可以省去每次买票排队的时间，以及恼人的语言障碍，何乐而不为呢？这似乎感觉一切都很划算。但实际上大部分人都无法把行程排得这么满，而常常会在迪士尼乐园玩上两天，或者到东京购物三天，导致实际上坐车的开销还不到票价的一半，这样 JRPASS 就不合适了，还不如每次都单独买票。其实这样的定价措施早已成为一种成熟的商业模式，铁路公司通过复杂的差别定价，一方面吸引需求各异的客户以增加收益，一方面间接促进当地旅游经济的发展，一举两得。

D1	成田机场-新宿	乘坐成田特快NEX	3200円≈¥200元
	新宿-东京周边	乘坐JR山手线前往新宿 上野/池袋/银座地区	900円≈¥56元
D2	富士山一日游	乘坐JR中央線通勤快速 河口湖行两个半小时 即可到达富士山	往返价格 4700円≈¥300元
D3	东京前往迪士尼	乘坐JR埼京线到新木场 换乘JR京叶线到舞浜	往返价格 1400円≈¥87元
D4	从东京前往大阪	乘坐新干线HIKARI号 或者KODAMA号	单程票价 14000円≈¥868元
	到达大阪后可前往难波附近		往返价格 480円≈¥30元
D5	可前往京都、奈良观光游览，大阪-京都 大阪 - 奈良		往返价格 ¥70元 往返价格 ¥100元
D6	可前往环球影城，体验哈利波特与好莱坞		往返价格 360円≈¥23元
D7	从大阪前往关西机场,乘坐关空特快HARUKA号		1200円≈¥75元

图 4.11　日本 7日火车票路段

这个差别定价原则可以借鉴，却不可以照搬照抄。我国的国情是，铁路的客户不是太少，而是太多了，特别是在春运期间。所以我们运用弹性的差异化定价原则，希望它达到的效果不是吸引更多的人，而是分散原来高度集聚的需求。例如，在春运期间，买票难也不是在各个线路各个方向同时成比例地难。一般来讲，上海到四川的车票难买，但是四川到上海的车票就相对容易。如果春运是因为大量离乡务工人员想回家造成的，而从沿海到内地的火车必须再开回沿海的话，那么几乎可确定：在上海通往四川和贵州的车票“一票难求”的时候，从四川和贵州开往上海的火车是“一客难求”。这说明现在的价格没有弹性，没有差异化，结果大量的运量没有释放出来，很多火车是在空跑。

如果能差异化定价，在提高从沿海往内地的车票价格的同时，大幅减少从内地向沿海的车票价格，会吸引部分民工不回家，把家里人接出来过年。这样一来，就可以部分缓解原来的“一票难求”。图 4.12 是大家熟悉的春运期间买票难的场景。



图 4.12　春运期间买票难

很多人说，这是不可能的，因为回老家过年是中国人的传统。但传统从来就不是一成不变的。过去上海人过年也回家，为什么现在很多上海人过年过到海南，甚至去东南亚了呢？再次，我们不是不承认这个传统的力量。弹

性定价，恰恰就是尊重了回家过年的传统，所以对于牺牲了回家过年的人，我们给予了奖励：你可以以很低的成本到上海和北京来看看。

我们还可以提前一年或者半年售卖春运车票，这有利于大家提前安排时间，也有利于铁路部门根据出售车票的情况安排运力。因为可以提前购票，也加大了黄牛的资金压力和风险，不利于黄牛的生存。提前购票还可以使得弹性票价成为可能。火车票可以向飞机票学习，在不同时间段给出不同的折扣。例如提前一年或半年购票的加价 100%，然后每个月下降 10%。这样对价格不太敏感而回家需求强烈的人可以早一点按照高价购买到车票，而对车票价格敏感的人可以等待并承担买不到的风险。这与目前廉价航空的越早订票越便宜恰恰相反，但原则是一样的，只是目的不同。廉价航空希望吸引客户，而铁路系统希望减少春运的需求。

最后，不能不面对的是一个补贴弱者的问题。上述经济手段调节需求的结果，简单言之，就是把票给了出价最高的人。如果一定要在春节的特殊时期，给收入较低的人回家过年以一定的福利，政府就需要给予财政补贴。但是，这种对待特殊群体的补贴票，很显然也应该实行实名制。事实上，现在给学生的折扣票就是一种实名制售票。不过这里涉及更复杂的身份认定问题，已经超出了一个公司的能力范围。更何况，中国的火车票市场是供小于求的问题，如果铁路总公司成为一个完全市场化运作的公司，这个问题就更难解决。有时候，再强大的算法也不得不面对现实社会的难题。

未来的动态定价——服务证券化

不同于传统算法，现在悄然主导我们生活的是“能够学习的机器”，它们通过学习我们琐碎的数据来执行任务；甚至我们还没提出要求，它们就能完

成我们想做的事。

——佩德·多明戈斯，《终极算法》

在商业社会高度发达，服务的供给和需求都在高频率地发生的背景下，我们不得不面对以下三个问题：

（1）如何快速撮合服务提供者和需求者，并给出公允的定价促使交易发生？

（2）当服务价格偏离公允价格时，系统如何修正不合理的定价？

（3）当服务行为发生在未来时，如何在当下进行定价和交易？

面对大规模、高速度的服务请求，现有的动态定价模式也越来越难以响应。在这里我们提出一种将服务作为一种有价证券，并投放二级市场交易（类似股票市场）的算法来解决这些问题。

公允价格的发现——集合竞价

为了更加形象地阐释服务证券化的理念，我们以顺风车服务的运营为例。假设小明是一个深圳的私家车司机，他第二天要去广州出差。小明想顺路做个生意，载客从深圳前往广州，于是他在市场上发布了一个卖单，希望卖出自己提供的载客服务，如图 4.13 所示。

卖出服务单
司机：小明
出发地：深圳
目的地：广州
时间：xxxx 年 xx 月 xx 日
要价：600 元

图 4.13　小明的卖出服务单

小李是一个深圳的学生，正好第二天想去广州游玩，于是他也在市场上发布了一个买单，希望买入自己需求的服务，如图 4.14 所示。

买入服务单 乘客：小李 出发地：深圳 目的地：广州 时间：xxxx 年 xx 月 xx 日 出价：300 元

图 4.14　小李的买入服务单

相似地，市场上有千千万万的司机发布了他们的卖出服务单，也有千千万万的潜在乘客发布了他们的买入服务单。市场运用大数据、人工智能等技术，对海量的买入单、卖出单的信息进行整合、匹配和处理。当卖出单提供的服务和买入单需求的服务相匹配时，系统进行集合竞价，直到买卖价格一致，交易达成。这个达成的交易，被像是一张证券，作为乘客在指定时间和地点，接受司机提供指定服务的凭证。

这种公开、透明、海量的交易，展现了一种发现公允价格的模式。人们可以通过参考交易系统公开披露的成交信息，对特定服务的价格有大致的评估。在网约车的应用场景中，油价的上涨、道路的维修等市场扰动会促使司机调整服务价格，乘客也会跟随卖出单给出新的报价。在这种买卖双方的博弈中，新的公允价格再次达成。

公允价格的守护——套取利差

证券化的交易系统是可以自我修正不合理的价格的，这一点需要依靠市场中的套利交易者。仍然以网约车服务为例，这里我们不考虑时间因素（图 4.15）。

假如一张从 A 地到 B 地的证券价格为 300 元，从 B 地到 C 地的证券价格为 500 元。而现在，一张 A 地到 C 地的证券价格为 1000 元。

显然，当 A-B 的证券和 B-C 的证券被公允定价时，A-C 的证券定价偏高了 200 元。此时，市场上的套利者发现机会，花费 800 元买入 A-B 的证券和 B-C 的证券，然后将两者合成为 A-C 的证券，再以 1000 元的价格卖出，净赚 200 元。

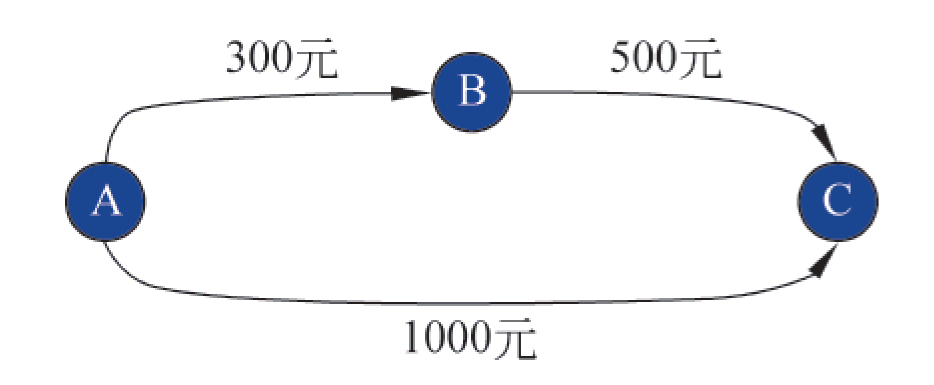


图 4.15　证券化市场的交易闭环

当套利者越来越多时，A-C 证券的卖单越来越多，价格下跌；A-B 证券和 B-C 证券的买单越来越多，价格上涨。这两个过程共同压缩套利空间，驱使三者的价格公允化，直到套利空间消失。因此，我们可以说套利是市场动态定价过程中，公允价格的守护者。

跨时间的资源整合——交易未来

服务证券化的另一个优势在于跨时间的资源整合。仍然以网约车服务为例，假定小王是一个深圳的私家车司机，上班之余偶尔为市场提供顺风车服务。这时候，小王的女朋友意外怀孕，突然要结婚，需要一大笔钱。此时，小王可以提前将自己未来 3 个月内提供的顺风车服务打包为证券，投放市场降价卖出，短时间内获得一大笔钱，解决燃眉之急。正如图 4.16 所示，在这个案例中，小王通过把未来的服务证券化，实现了“向未来的自己借钱”的诉求。

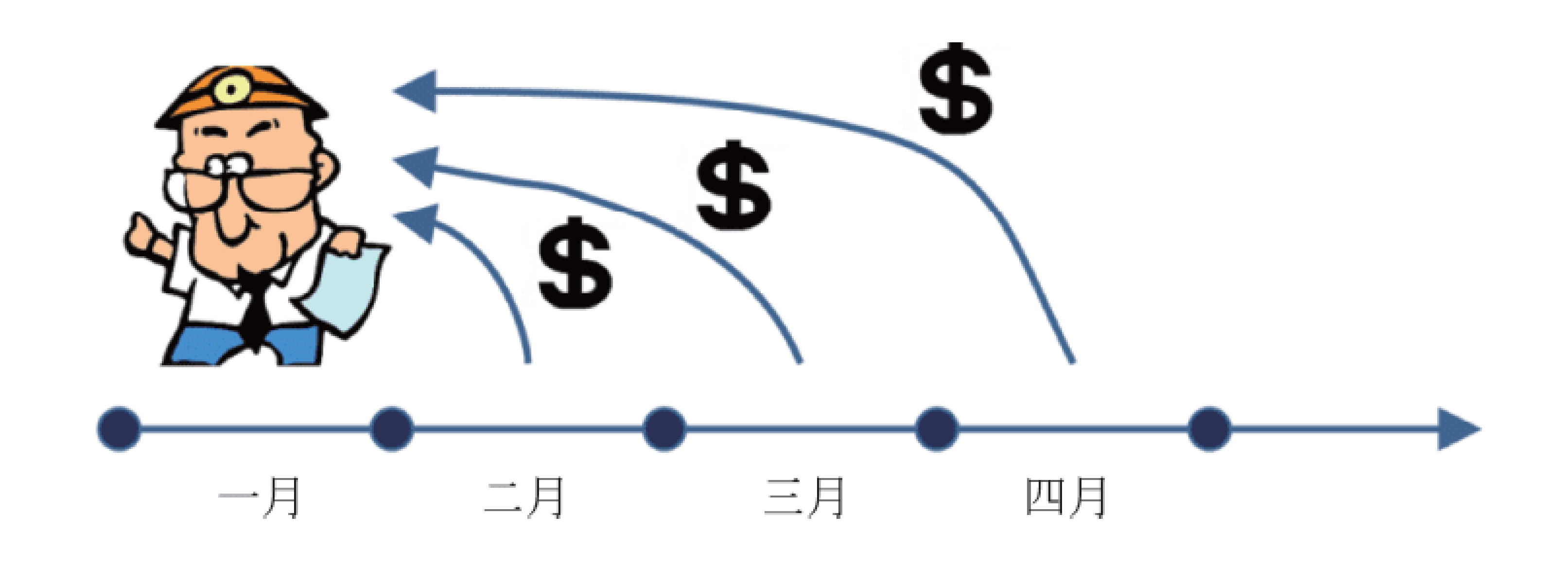


图 4.16 证券化折现未来服务

现在我们从服务购买者的角度，考虑另一种情形。在深圳工作的小方最近想去广州探望父母，于是他在市场上购买了小王出售的证券，预期下个月坐小王的车去广州。结果就在出发的前一天，小方的老板要求小方去南非出差半年。小方赶紧把自己手头的证券重新投放市场，降价出售。在这种意外发生的情形下，乘客小方通过证券交易避免了更多的资金损失，司机小王也并没有因为小方的突发事务失去生意，是一个双赢的结局。

无论是司机小王的案例，还是乘客小方的案例，他们都通过降价出售证券来满足自己各自的需求。仔细思考可以发现，这里“降价的部分”是出价者对“快速响应要求”的定价。例如，小王要求快速拿到未来的钱，以降价的方式加快出售；小方要求快速出手证券，以免服务过期，以降价来吸引市场购买。

因此我们可以发现，服务的证券化，使得对未来进行交易成为可能，也提高了服务行为的效率。

算法重新定义价格

在赛博新经济时代，互联网成为现代社会的基础设施，而网络对我们生活最深刻的影响就在于“留痕”，留下的痕迹就是数据。数据之所以产生价

值，不在于“量大”，而在于“线上”，因为数据从此可以在更大范围内流动，进而产生更有影响力的价值，这是真正的数据带来的巨大变化。**阿里巴巴技术委员会主席王坚的著作《在线》中也提到，当今社会，“互联网”即“基础设施”，“数据”即“生产资料”，“计算”即“公共服务”，三者聚合在一起，就是“计算经济”。计算经济中很重要的一环就是发现并确定数据价值**。数据的价值是靠计算发现的，云计算以公共服务的形式让每个人都能获得 AlphaGo 所需的计算能力，计算资源容易获取，不再被垄断。正因为如此，在赛博新经济时代，通过机器学习、数据挖掘等计算方法处理在线数据，用更高效的手段重新发现商品价格，才能更理性地指导人们在赛博时代下的经济生活，从而建立适应时代变化的市场新秩序。

对于动态价格的制定规则，考虑到成本、技术、用户接受程度、社会影响等诸多要素，目前，大部分的应用场景还集中在航空、酒店、票务、电商、出行等领域。一方面，由于这些必须消费品时刻伴随着我们的生活，从本地服务到外出旅行；另一方面，用户在与这类产品长期的磨合和适应过程中，已经养成了习惯动态定价的消费观念。

事物总有它的两面性，动态定价虽然有助于发现市场价格，却也带来了诸多问题，例如对用户不友好、影响客户忠诚度，以及企业口碑等。以 Uber 为例，可能更多用户不会理解动态定价是为了平衡供需，而是会觉得 Uber 在急用车时翻倍涨价太唯利是图、赚黑心钱。再以票务为例，如果在采用动态定价时，出现了两张邻座票价格不同的情况，买高价票的就会觉得自己被欺骗了，从而对售票方表示不满。尤其是高铁火车票这类涉及普通群众切身利益的商品，一旦通过定价调整供需有失偏颇，不仅关系到公平性，更会对社会制度产生不利影响。

如何降低这种影响是采用动态定价的企业都需要认真考虑的问题。如果考虑的因素不周全，那么对顾客的感知价格评估就不准确，动态定价的意

义对于传统定价的优势也就越小。如果算法没覆盖到，一旦遇到突发事件，反而可能带来巨大的负面影响。此外，动态定价对技术要求比较高，要考虑投资回报率。对于某些企业而言，特别是传统企业，订制一套动态定价系统的成本并不低，相比它可以带来的潜在收益，值不值得投入还得根据企业自身的需求进行评估。所以，并不是每个企业都有能力采用动态定价。

而采取创新的运营策略也可以积极降低这种影响，同时给用户带来良好的使用体验。例如滴滴的“拼车”功能，就是根据交易撮合的程度进行动态定价。当乘客计划拼车出行前往某个目的地，他在下单时会选择“目的地”和“拼车人数”，平台便会给出“一口价”，该价格在进入行程后不再改变。例如，从北京南站到首都国际机场的里程费预计 100 元，平台加上服务佣金预估价 160 元，当有 4 人选择拼车同行，平台给每人出价 40 元，这样，乘客便以较低的价格共享了该行程，同时司机和平台都赚取了一定的费用。但是这样双赢的局面不会常常出现，为了提高成单率，有时候，即便拼不满 4 个人，系统也会分配司机接单，假如只有两人参与以上行程，平台就需要补贴 20 元，这样的情况多了，滴滴难免亏损严重。

那么，如果根据交易撮合的程度进行动态定价，当拼车人数不足以覆盖成本时，系统给出“两口价”，即拼单成功一个价格，不成功是一个价格，用户可以权衡时间和价格后，选择自己更能够接受的行程，系统也不至于亏损，一些城市正在推广实行该策略。如图 4.17 所示，2017 年 5 月 26 日，滴滴在北京推出了全城拼车体验活动，乘客在固定站点上车，享受“一口价”服务。



图 4.17 滴滴“一口价”拼车活动

新经济模式对传统定价方法提出的挑战越来越多，趋向高频交易特征的需求变化速度日益加快，对于某种服务的买卖双方，供给该如何适应需求的高速而剧烈的变化，在不能适应的状态下，又该如何找到强需求的用户，并且精准营销进而提供专业服务，或许只有足够智慧的算法通过重新定义价格，才能更好地将这些问题解决。

随着每个细分领域技术的不断迭代，以及对用户需求更加精准的把握，基于算法的动态定价将以更细微的粒度、更敏锐的时间、更合理的价格变革传统的定价方式。类似股票市场中的高频交易，每一笔价格会在反复的博弈过程中逐渐趋于理性，直到达到市场的均衡状态，在最后一飞秒内成交。那时，全新的商业模式将为动态定价带来再一次的颠覆性革命。

第5章　算法，让数据有了价值

今天的人们早已从信息匮乏的时代进入了数据大爆炸的时代，每个人都是信息消费者，也是信息制造者。然而，我们却越来越难以从大量信息中找到自己感兴趣的内容，也越来越难以让自己生产的信息脱颖而出，被别人看到。在数据的汪洋大海中，并不是每一朵浪花都具有价值，可我们却常常被那些无用的信息裹挟，深陷其中，迷失方向。数据本身并不产生价值，按照大数据专家的观点，首先需要挖掘数据中的信息，然后从信息中建立结构并发现知识，再把知识转换成能力，基于能力建立系统，系统中策略的执行才最终产生价值。这个链条可真长！我们可以简单理解为只有有序的数据，能够为用户提供有用信息的数据才能产生价值。那么到底如何才能从数据中找到价值呢？在赛博的世界里，再也没有谁比算法更有能力给数据赋予价值了。算法是如何做到这一点的呢？

大数据的发展带来了人类新文明

计算机总是越来越智能的。科学家告诉我们不久它们就能跟我们对话了(这里的“它们”，我指的是“计算机”。我怀疑科学家永远都不能跟我们对话)。

——达沃·巴里，作家

每个时代都有自己的“大数据”

如果我们回顾历史，会发现每一种文明的更替都伴随着信息载体的变迁，而信息载体的变迁，又带来了信息量的增长。无论在哪个时代，都存在着对于那个时代的信息载体和信息处理工具来说是“规模很大”或“很复杂”的数据集合。从人类诞生之日起，数据就不断产生，例如部落的人口、捕杀的猎物、农田的收成，甚至口口相传的传说都可以视为数据。数据逐渐产生、积累，导致数据规模越来越大，数据之间的关系也越来越复杂。当数据规模和复杂程度逐渐达到甚至超出人类当时的数据处理能力后，即可被视为“大数据”——尽管当时的人们并没有“大数据”的概念。

在文字诞生之前，原始先民生活中发生的事件只能靠个人头脑记忆。后来发生的事情越来越多(即“数据”规模越来越大)，只靠记忆难免发生混乱。为了避免忘记以往的事情，先民们需要花上一整天的时间来复习过去发生的事情，然而悲催的是，每天都有新的事件发生。这无疑使得原本困难的生活雪上加霜，记忆和回忆占用了大量时间，都没有足够的时间进行渔猎填饱肚

子了！

幸运的是，传说中神农氏发明了结绳记事，拯救万民于水火，使得先民不必每天花费大量精力记忆众多的琐事。《周易·系辞下传》中说："上古结绳而治，后世圣人易之以书契。"其意思就是，上古的人们通过在绳子上打结来记事，后世的圣人发明了文字，开始用书写和文字来代替结绳记事。东汉经学家郑玄在所著的《周易注》中解释说："古者无文字，结绳为约，事大，大结其绳；事小，小结其绳。"《周易集解》进一步解释说："古本无文字，其有约誓之事，事大，大其绳，事小，小其绳，结之多少，随物众寡，各执以相考，亦足以相治也。"图 5.1 给出了用于记事的绳结示意图。

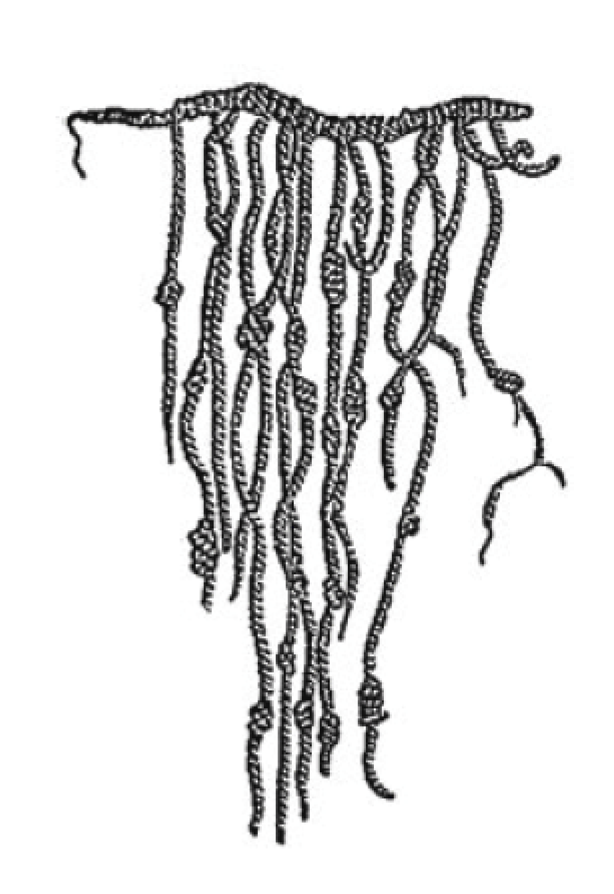


图 5.1　用于记事的绳结

结绳记事虽然方法简单，实际上也包含一套相对完备的"算法"，即大事对应大绳结，小事对应小绳结，数量多就多打绳结，数量少就少打绳结，这套简单的机制在原始社会中"足以相治"，可以满足一般的生活需求。结绳记事这种新的"数据处理"方式，配合语言交流，使人们可以更加方便地记住更多的事情，应对了那个时代的"大数据"挑战。与此类似的还有农耕文明中的甲骨文和封建文明的造纸术。

近年来，随着互联网和移动通信技术的发展，全球数据总量呈爆炸性增长，增长速度也逐年加快。1984 年诞生于美国旧金山的思科(Cisco)公司，是全球 IT 企业中的巨无霸。作为全球顶级的网络解决方案提供商，仅 2016 财年第四季度的收入就达 126 亿美元。思科公司在《思科可视化网络指数：全球移动数据流量预测白皮书(2015—2020)》中，对 2015 年全球的移动数据流量做了详细的统计分析，并对未来五年的移动数据流量的增长做出了预测，如图 5.2 所示。与 2014 年相比，2015 年全球移动数据流量增长了 74%。另

外一个令人震惊的数值是，2015 年全球移动数据流量是 15 年前的 4 亿倍，是 10 年前的 4000 倍。截至 2015 年底，全球每个月产生的移动数据量平均为 3.7EB，而在 2014 年底时，这个数值是 2.1EB。据思科公司预计，到了 2020 年，全球月均移动数据流量将增长 8 倍，达到 30.6EB。这个数据量堪称惊人。要知道，1EB 数据包含 1 152 921 504 606 846 976 个字节，存储 1EB 的数据需要 1 048 576 块 1TB 的硬盘。

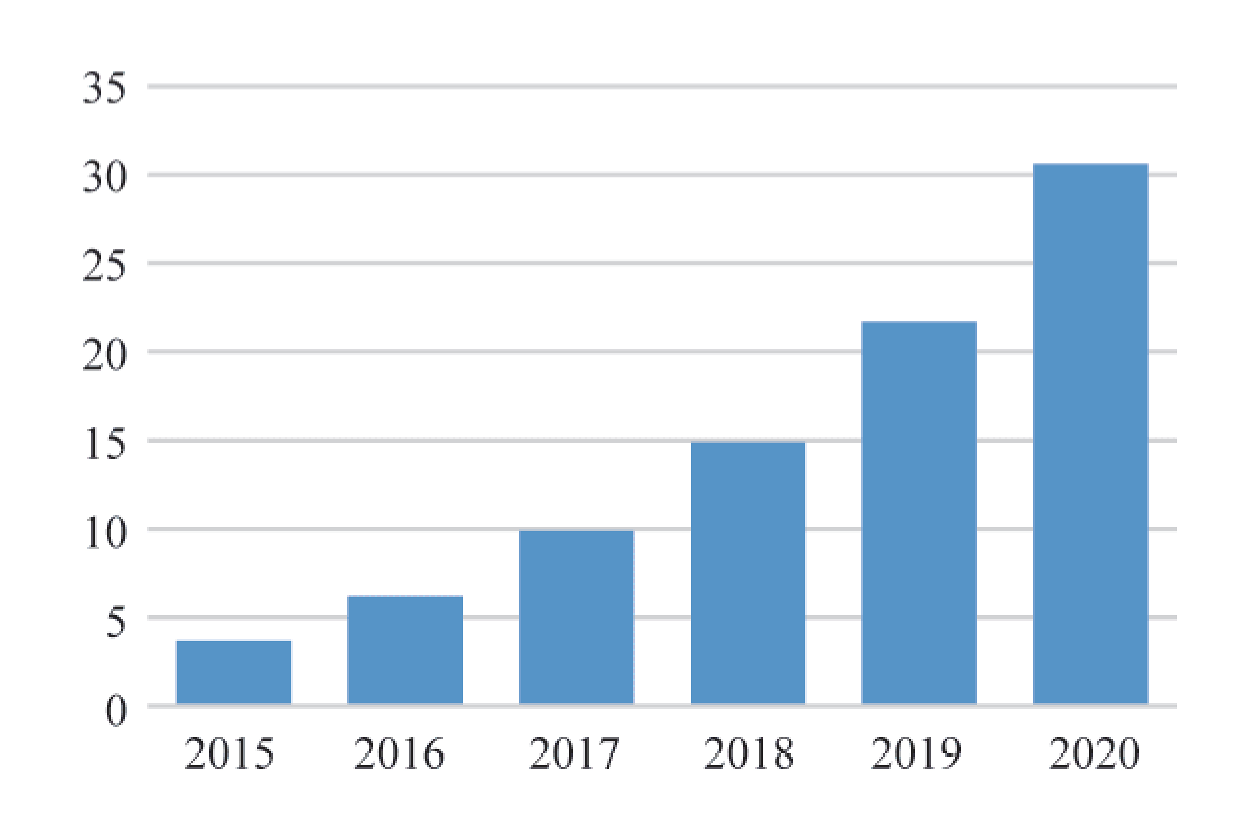


图 5.2　思科公司预测 2015—2020 年间全球移动数据流量增长情况（单位：EB）

全球著名的科技咨询顾问提供商国际数据公司（International Data Corporation，IDC），服务领域主要集中在信息技术和电信等行业。IDC 在报告 *THE DIGITAL UNIVERSE IN* 2020：*Big Data，Bigger Digital Shadows，and Biggest Growth in the Far East* 中指出，2020 年的全球数据总量将是 2005 年的 300 倍，达到 40 000EB；并且从 2013 年开始，全球数据总量将 2 年左右就翻一番。根据专家的预测，到 2020 年，每位互联网用户平均每天要产生 1.5GB 的数据流量，一辆无人驾驶汽车每天产生的数据量将达到 4000GB，一架飞机每天产生的数据量将达到 40 000GB，一座智慧工厂每天产生的数据量更是超过 1 000 000GB。

大数据，实际上就是人类文明的产物。不同的时代，因为数据处理能力不同，对何种数据集合是“大数据”也有不同的认识。过去的大数据，可能就

是现在的“小数据”;现在的大数据,可能就是将来的“小数据”。大数据永远是一个相对的概念。

运转不停的世界每天都有大量数据产生,未来还将有更加大量的数据产生。并且,在信息技术的催动下,数据洪流以前所未有的速度席卷到我们生活中的每个角落,我们每个人都身处其中,没有人能够例外。

赛博时代的大数据

虽然每个时代都有自己的大数据,但是大数据真正被我们熟知却并没有多久。一般认为,20 世纪 90 年代,由于计算机科学家约翰 · 马西(John Mashey)的大力推广,大数据一词才慢慢流行起来。我们能从图 5.3 的百度指数清晰地看出 2011—2016 年以“大数据”为关键字的搜索趋势(上面曲线)和媒体关注(下面曲线)变化情况。图 5.4 是谷歌趋势给出的 2004—2016 年以“Big Data”为关键字的搜索热度。从这两个图都可以看出,大数据一词大概于 2012 年才慢慢进入公众视野,并成为一个越来越热门的话题。

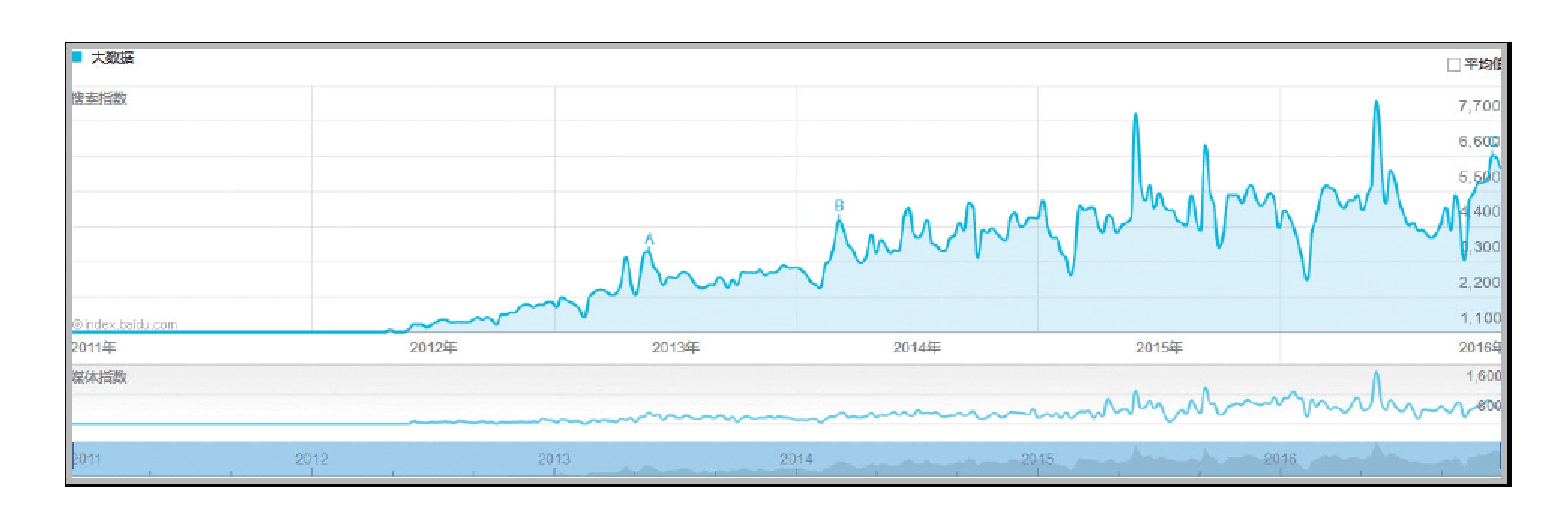


图 5.3 “大数据”一词的搜索指数和媒体指数

根据维基百科的定义,大数据(Big Data),又称巨量数据、大资料,指的是所涉及的数据量规模巨大到无法通过人工或计算机在合理的时间内达到截取、管理、处理并整理成为人类所能解读的形式的信息。百度百科则将大

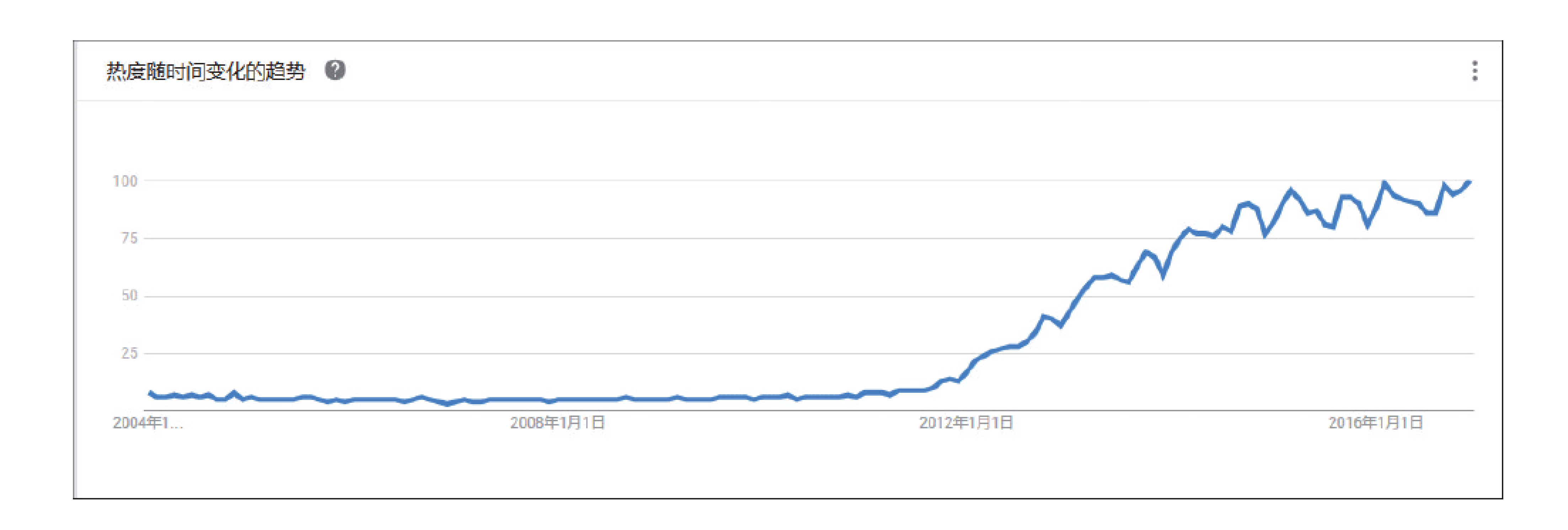


图 5.4　"Big Data"一词的搜索热度

数据定义为：无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

这听起来也许有些抽象。如果不使用这些严谨的科学术语，大数据到底是什么呢？大数据和普通的数据有什么区别呢？其实，通俗地说，如果数据规模不断增长或者复杂程度不断提升，以至让我们觉得这些数据处理起来很棘手时，普通的数据就一跃成为"大数据"。

大数据一词流行之前，人们通常使用海量数据（Mass Data 或 Massive Data）一词来指代大规模的数据集合。但是，我们不应简单地把"大数据"当作"海量数据"的同义词。事实上，赛博时代的大数据除了数据规模大之外，通常还具有流转速度快、数据类型多、价值密度低等特点，这就与以往"海量数据"有了本质区别。

大数据蕴含的大能量

人类活动产生了大量的数据，这些大数据中存储了大量的信息，大数据对人类的价值就在于这些信息的价值。我们先从一件小事中一窥大数据蕴

含的“能量”。

1986 年，美国自动化领域的先驱、迪堡集团有限公司（The Diebold Group，Inc.）的创始人约翰·迪堡（John Diebold）提到了这样一件趣事。1979 年，一家银行在某地区安装了 ATM 机网络。但是银行发现，某一台取款机每天凌晨 0 点到 2 点之间都会发生大量的提款操作。这是怎么回事呢？银行怀疑这里发生了违规行为，于是专门雇用了侦探对此进行调查。调查的结果充满戏剧性，侦探发现这个提款机刚好位于红灯区附近，午夜光顾红灯区的人们当然不想刷卡消费，因为这会在信用卡交易中产生一笔“不光彩”的记录。于是，这些顾客总是先在这部 ATM 机上取出现金，然后去红灯区消费。

如果通过一台提款机的取款记录能算出某个特定区域的消费特征，那么如果把全国所有提款机的取款记录都做采集分析，再辅以特定的计算机算法，能算出的就不仅限于红灯区消费了，甚至可以更详细地了解全国人口的储蓄能力、消费水平、消费习惯、资金往来、资产负债等情况。

大数据还可以让搜索引擎在公共卫生领域发挥作用。谷歌公司曾于 2008 年推出了流感趋势系统（Google Flu Trends）和登革热趋势系统（Google Dengue Trends），这套系统采集了美国人平时经常使用的 5000 万个搜索关键字和 2003—2008 年间美国疾控中心公布的疾病传播数据，发现了特定搜索关键字和疾病传播数据之间具有相关性。如果某地区包含特征字段的搜索关键字数量有异常上升，就可以认为该地区已经开始发生疫情传播。

为了充分挖掘大数据中存储的有效信息，谷歌公司共测试了 4.5 亿个数学模型，并从中确定最有效的预测模型。通常，美国疾控中心（Centers for Disease Control and Prevention，CDC）公布的疫情传播报告通常比实际传播情况滞后两周。因此谷歌公司声称，这套系统可以比 CDC 更早获知流感疫

情，并且该系统预报的疫情和官方公布的数据相关性高达 97%。虽然也有专家对此数据表示质疑，但不可否认的是，大数据技术正在以我们意料不到的速度和方式渗透到我们的生活里，为研究人类行为和人与人之间大规模的互动提供了崭新的思路。

2015 年 1 月 25 日，朋友圈发生了一桩“灵异”事件——一些人发现“宝马中国”在他们的朋友圈发了一条状态；一些人的朋友圈出现了“vivo 智能手机”的文字和图片，而另有一些人的朋友圈则出现了“可口可乐”。很快真相大白，原来在这一天，已经积攒了 5.49 亿用户的微信，正式启动了基于朋友圈大数据分析的新型广告推送(图 5.5)。



图 5.5　微信朋友圈的第一条广告

与传统广告全面铺开的传播方式相比，微信试图通过用户大数据分析，把广告推送给可能会感兴趣的用户。通过用户的注册信息、绑定的手机号和 GPS 定位，可以判断出用户所在地，通过朋友圈里分享的帖子，则可以分析出用户的职业特点、兴趣爱好等，钱包支付记录则体现出用户的消费水平，卡包里各种会员卡体现出用户的品牌倾向，除此之外，在朋友圈的点赞、留言也都可以体现出用户的个性和偏好等。这些庞大的用户行为数据如同一座宝藏，微信可以通过不同的算法了解到用户的特点和需求，从而推算出用户感兴趣的东西。

算法，点石成金的力量

你把数据拷问到一定程度时，它自然就会坦白一切。

——罗纳德·科斯，诺贝尔经济学奖获得者

大数据并不等于大价值，就像开采矿藏一样，没经过开采的数据仅仅是一堆数据。而借助算法，才能挖掘出大数据中熠熠生辉的珍宝。算法，赋予了大数据生命，让数据可以为我们所用。

机器学习算法——计算机会学习

随着数据规模的增大和复杂程度的提升，人类自身的力量变得越来越渺小，如此规模的数据显然是人力难以处理的，所以只能求助于计算机和数据处理算法。实际上，在大数据面前，简单的算法也往往因为效率太低而难以奏效，于是人们寄希望于计算机可以自己学习大数据中存储的信息，这就是机器学习。

从“机器学习”这个名字不难看出，这门科学研究的就是如何使机器具备学习能力，进而掌握知识具备智能。可以说，机器学习的终极目标就是人工智能。机器学习的产生和发展与人工智能技术的发展息息相关，其应用范围也主要集中在人工智能领域。

这两年人工智能步入了蓬勃发展的阶段。但是，人工智能从诞生到现在，也经历过三起三落，既获得过人们热切的期盼，也经历过漫长的寒冬。人工智能的历史，是一部群星璀璨、高潮迭起的历史。

作为一门科学，人工智能诞生于 1956 年的达特茅斯（Dartmouth）会议。这次会议上，相关领域的科学家就这门学科的名称达成一致，即 Artificial Intelligence（AI），并且明确了人工智能研究的任务。

人工智能自诞生以来，大致经历了三个阶段，分别是推理期、知识期和学习期。

从 20 世纪 50 年代中期到 70 年代初期，可以称为人工智能的“推理期”。由于达特茅斯会议的影响，这一阶段人工智能得到了大量资金投入，受到了业界、政府和学术界的广泛关注，这是人工智能发展的黄金时期。然而，就如同大多数新技术成长过程往往充满曲折一样，人工智能也是如此，人们慢慢发现他们当初的乐观预期迟迟无法实现。于是，20 世纪 60 年代后期，人工智能研究渐渐步入低谷。

20 世纪 70 年代中后期到 80 年代中期，是人工智能发展的第二个阶段——“知识期”。当时，以爱德华·费根鲍姆（Edward Feigenbaum）为代表的科学家提出，让机器拥有智能的前提是使机器拥有知识。于是，“知识处理”和“知识工程”成为人工智能的主流研究方向，“专家系统”（Expert System）开始被大家接受。专家系统是具备专业知识的计算机智能程序，可以扮演领域专家的角色。专家系统带动了人工智能的再次繁荣。1981 年，日本启动了以面向知识处理为主要目标的“第五代计算机”计划。然而，随着研究深入，人们发现知识的表示是一件很困难的事。于是，知识工程进入瓶颈期，“第五代计算机”预先设定的目标也没能按期实现。与此同时，由于世界范围内经济泡沫破裂，人工智能再次进入低潮，进入了人工智能的“寒冬期”。

20 世纪 90 年代后期，人工智能进入“学习期”。知识的表示是一个难题，那么能不能让机器自己学习知识呢？在这个阶段，机器学习融合了符号主义和联结主义，成为人工智能研究主流，产生了很多让人眼前一亮的成果。目前，机器学习已经成功应用到计算机视觉、自然语言处理、搜索引擎、语音

和手势识别等诸多领域。

尽管人工智能步入学习期的时间还不长，但使机器具备学习能力却是人类一直以来的梦想。在人工智能诞生之前，就有人对机器学习进行了技术探索，其中的一位代表性人物就是香农。

1916年4月30日，现代科学史上最为独特的科学家——香农诞生于美国密歇根州。香农性格乐观开朗，年轻时曾有过骑着独轮车耍着四个瓶子在贝尔实验室招摇过市的霸气之举。香农一生学术涉猎非常广泛，本科学习数学，硕士期间研究布尔代数和电子电气，博士期间研究生物遗传学和代数学，二战时推进了密码通信的进步，二战后创立了一门崭新的科学——信息论。

二战期间，香农是为战争服务的科学家小组的一员。当时，他主要负责的是盟国之间的保密通信，还曾保障过罗斯福与丘吉尔的越洋通话。与香农一起为战争服务的还有“图灵机”模型的提出者图灵，图灵的主要工作是破解德国的通信密码。尽管工作内容有一定的相似性，但是由于严格的保密规定，两人相互并不知道对方的研究内容。虽然不能透露工作内容，但是两人经常在午餐时间共同讨论一些有意思的话题。有一次他们聊到了 Thinking Machine 的话题，就是让机器具备思考能力。这个话题刺激了香农的创新神经，于是他考虑真正实现一个具备思考和学习能力的机器，这就是著名的“老鼠迷宫”(图5.6)。

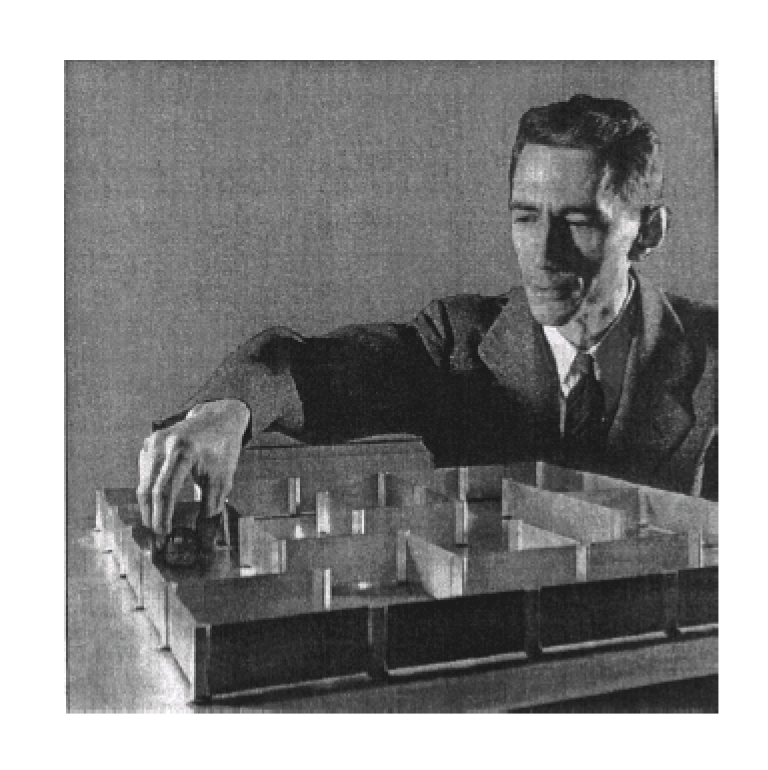


图5.6　香农和他的迷宫

迷宫的主体是5行5列的方格阵列，每个方格中都有一个可以固定目标的插孔，方格之间用可以插拔的“墙壁”隔开。迷宫中的目标可以用针固定在迷宫中的任何一个方格中。迷宫的主角是一只四处嗅探的“老鼠”。老鼠由两个电机驱动，一个电机负责前后移动，一个电机负责左右移动。老鼠在方

格中行走，寻找它的目标。寻找过程中老鼠可以感知到墙壁的存在，碰到南墙就回头。

这只老鼠具备三种能力，一是记忆，二是寻找，三是忘却。

所谓“记忆”，是指这只老鼠可以记住之前走过的路，找到目标后，还会把整条路线记在脑海里。当然，老鼠是通过大量继电器的开关状态实现记忆功能的，一旦断电记忆就会清零。老鼠行走到一个方格中心时，会根据“记忆”中的路线决定行进方向，决策机制如下：

如果目标就在当前的格子 A 中，老鼠就停止行进，点亮一盏信号灯，接通警报器，告诉它的主人香农：我找到目标啦！向主人卖萌的同时，它还会把起点到目标的路线记下来，这可是它以后偷懒的本钱！如果把老鼠再次放到起点方格中，老鼠会很快沿着相同路线轻松找到目标。

如果当前格子 A 中没有目标，但老鼠发现这个格子是记忆中“偷懒路线”的一环，那么它就果断开启偷懒模式，试图通过偷懒路线到达目标。

如果目标不在当前格子 A 中，并且格子 A 也不是“偷懒路线”的一环，或者老鼠记忆中根本就还没有所谓的“偷懒路线”，那么老鼠就只能采取最笨的方法了。它选择一个方向做试探，如果碰了壁就退回方格中央，逆时针旋转 90°继续尝试，直到顺利走到下一个方格 B。对于走过的每一个方格，老鼠会记下它离开这个方格时的方向——前、后、左或右。最坏的情况是，老鼠在这个方格的所有方向——当然，除了它进入这个方格的方向——都碰了壁，那它就只能退回上一个格子继续探索。

在这个策略的驱动下，老鼠通常会找到一条到达目标的路，但是在特定情况下，老鼠也会陷入某种“神经质”的状态。例如在老鼠记忆中，从格子 A 出发，经过格子 B 和格子 C 可以到达目标所在的格子 D。如果把迷宫稍做修改，使得从 C 出发无法到达 D，但可以到达格子 E，而经过格子 E 又恰好可以到达格子 A。这时，老鼠就会进入 A-B-C-E-A-B-C-E 的循环中。

如何解决这个问题呢？这是一只擅长偷懒的老鼠，所幸它也是一只聪明的老鼠。如果在24次移动后还没有到达目标，老鼠会认为迷宫结构已经发生了变化，它会果断忘掉曾经的“偷懒路线”，放空心灵，重新开始迷宫冒险之旅。

图5.7是迷宫的设计图，图中右下角是香农名字的缩写“C. E. S.”，中间部分是迷宫的主要部件——25个方格，上方是负责老鼠左右移动的电机，左侧是负责老鼠前后移动的电机，下方是由按钮、开关和拨片组成的控制面板，右上角的同心圆代表老鼠寻找的目标。

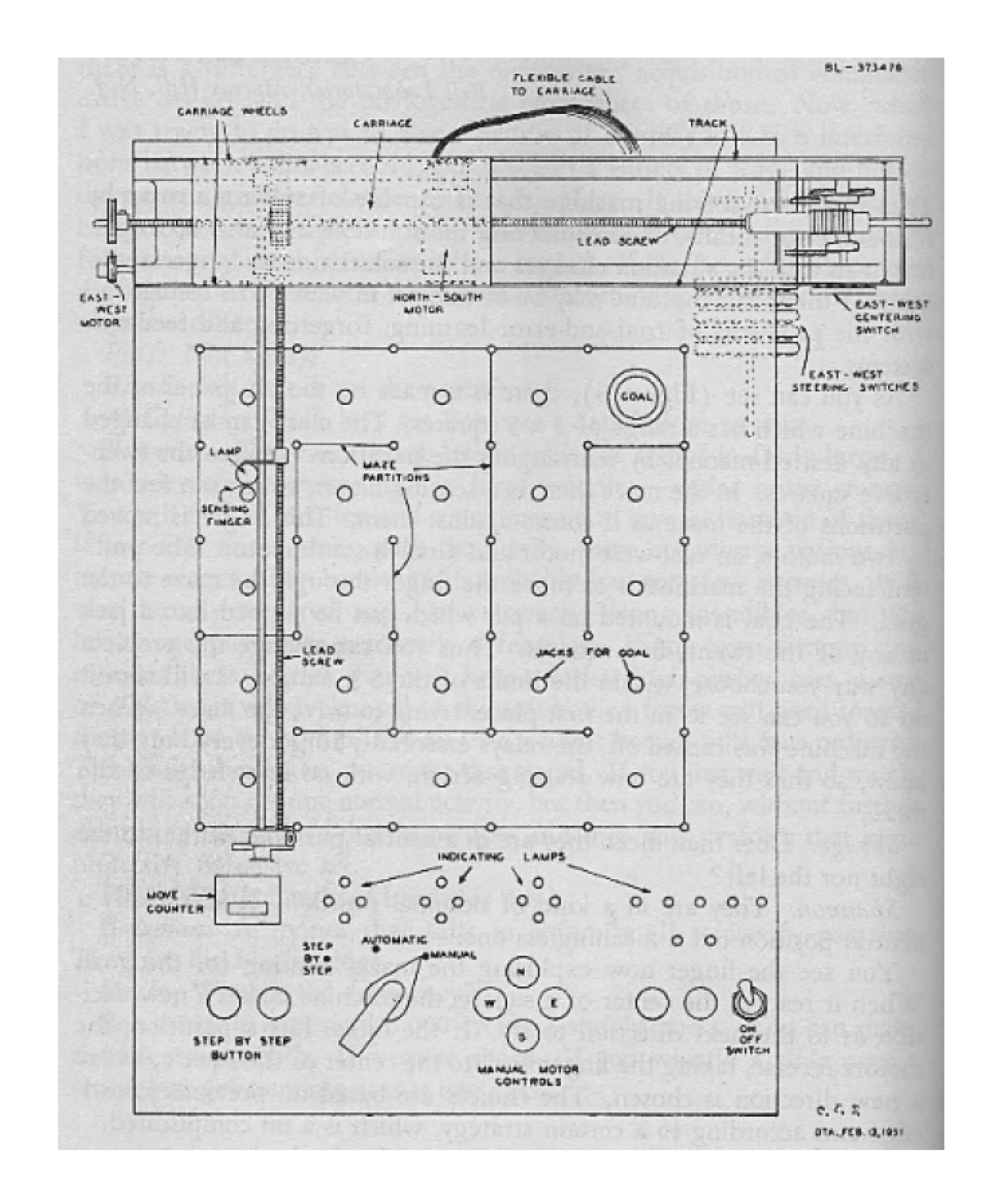


图5.7　迷宫结构图

从现在的眼光来看，这个迷宫功能比较简单，老鼠的破解能力和学习能

力也比较有限，甚至不能保证可以找到到达目标的最短路线。可是不要忘了，这个迷宫诞生在电子管和继电器的时代，那时现代意义的计算机还没有诞生，C 语言等编程语言也还没有发明。香农创造的这个迷宫具备了“记忆”“寻找”和“忘却”能力，老鼠的大脑就是一个简单的“算法”。这个迷宫体现了香农卓越的学术水平，以及他在使机器具备学习能力方向所做的思考。

机器学习是一门涉及统计学、概率论、最优化理论、计算复杂性理论等多个学科的综合学科。目前，机器学习并没有统一的定义。人工智能先驱亚瑟・塞穆尔（Arthur Samuel）认为，机器学习的研究目的是使计算机在没有被显式编程的前提下具备学习能力。另外一个更加常用的定义是由机器学习专家、美国工程院院士汤姆・米切尔（Tom Mitchell）给出的：如果计算机程序借助以往的经验可以使自己的性能变得更好，这时我们就认为这个程序具备了学习能力。

谈起机器学习时，我们要区分两个概念：机器学习算法和机器学习。机器学习算法，是使计算机具备学习能力的算法。机器学习，是设计和分析机器学习算法的学科。

大体来说，机器学习算法可以分为有监督算法（Supervised learning）与无监督算法（Unsupervised learning）两大类。两类算法的主要区别在于，作为输入的数据是否带有标签（Label）。若要求所有数据都带有标签，那就是有监督学习；若所有数据均不带标签，就是无监督学习。

什么是数据的标签呢？简单地说，就是对于数据性质的说明。以国内某音乐软件为例，对于听到的每一首歌曲，用户都可以做出收藏、删除、下一首等操作。这样，软件就可以根据用户操作为音乐生成一个标签：喜欢（单击“收藏”按钮）、一般（听完歌曲而不做任何操作）、不太喜欢（单击“下一首”按钮）或者讨厌（单击“删除”按钮）。后台运行的机器学习算法可以对带有标签的数据进行学习，明确用户的偏好，精确推送用户可能喜欢的歌曲。这就是

一个典型的有监督学习的例子(图 5.8)。

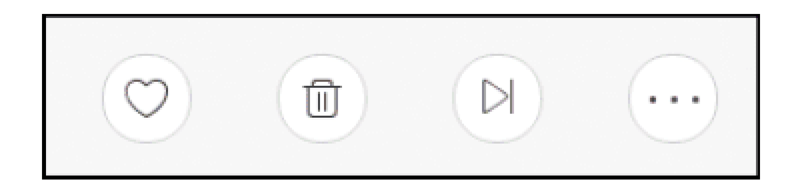


图 5.8　某音乐软件界面(从左到右分别是“收藏”“删除”“下一首”和“其他功能”)

在医学领域,医生们遇到疑难杂症时,常常阅读相关的文献来寻找以往的类似病例和治疗方案,为解决临床问题提供思路,这个过程称为系统性回顾(Systematic Review)。在系统性回顾中,最耗时、最乏味同时也最重要的一个步骤就是文献筛选,就是从大量文献中筛选出真正有用的部分。通常,每 2000～5000 份文献中只有 200～500 份文献最为相关,比例不足 10%。以往,文献筛选通常由人工完成,往往会耗费大量的人力和时间。

美国社会保障局(United States Social Security Administration)曾经主导了一个医学项目,旨在研究残疾儿童中低出生体重、发育不良和身材矮小之间的关联。这个项目交给塔夫茨大学的循证医学实践中心(Tufts Evidence-based Practice Center),他们需要筛选 33 000 篇文献。一个资深的文献筛选员可以在一分钟之内阅读两篇文献的摘要并进行筛选,这意味着筛选 5000 篇文献需要 40 个小时的不间断工作;而某些晦涩难懂的文献可能需要数分钟才能完成筛选,这会使得所需时间大大增长。

为了解决这个问题,塔夫茨大学计算机科学系的研究人员提出,使用一个有监督的机器学习算法完成文献筛选工作。首先,请医学专家筛选出数篇文献,这相当于对一部分文献加了标签,“相关”或“不相关”;然后由机器学习算法对筛选好的文献进行学习,并对剩下未筛选的文献进行预测分类。经测试,人类专家只需对 50 篇文献做筛选,机器学习算法就可以达到 93% 的筛选准确率,并且筛选工作可以在一天之内完成。

当然，数据量过大时，为所有数据增加标签是不可行的，于是无监督学习算法应运而生。无监督学习算法一个常用的场合是为数据分类。例如，搜索引擎的新闻分类显示功能，后台运行的可能就是一个无监督学习算法。每天世界上都会发生海量的新闻事件，这些新闻不可能由人手动设置标签。无监督的机器学习算法可以自动对新闻进行学习、分类。谷歌公司在其网站上注明："所有新闻的选择、排序、分类和搜索均由电脑程序自动决定"（见图 5.9）。



图 5.9　谷歌新闻界面

1996 年，美国数学家、物理学家、计算机科学家大卫·沃尔珀特（David Wolpert）提出了一个著名定理——"没有免费的午餐"定理（No Free Lunch Theorem）。这条定理表明，一个算法不可能在所有问题上都表现得很好，算法在解决某类问题时性能的提升，是以在解决其他问题时性能下降为代价的，不存在满足所有需求的"万金油"。这条定理给机器学习的启示是，没有一个机器学习算法可以擅长学习所有的模式。对于具体问题，还需要具体分析，设计出适合的机器学习算法。虽然如此，我们仍然可以期待技术的发展为我们带来相对通用的算法组件，通过这些组件可以组合、训练出具有特定用途的机器学习算法。

搜索算法——如何海底寻针

20 世纪 90 年代以前，互联网的用户主要是军方和科研机构，网络上的信息也很少。所以，这个阶段人们并没有什么搜索信息的需求。这就好比去一家便利店买东西，因为店面很小，货物很少，所有货物一目了然，我们很容易找到自己需要的东西。1992 年，万维网的出现打破了这一局面，网站制作和发布网页信息的成本急剧下降，普通用户浏览信息的成本也急剧降低，互联网上的信息开始快速增长。在信息快速增长的情况下，用户想找到自己要看的网页内容变得越来越困难。

1994 年，斯坦福的三名在校生建立了一个名为雅虎（Yahoo）的网站，他们自己将互联网上重要的站点分门别类整理成一个导航目录，并把这个导航目录放到雅虎网站上，方便用户快速找到希望访问的网站（图 5.10）。雅虎一经推出就获得了市场的认可，用户渐渐形成了首先访问雅虎然后再访问其他网页的习惯。这就好比去一家沃尔玛超市买东西，店面很大，货物很多，我们必须通过分类指示牌才能找到所需货物摆放的货架，雅虎正是起到了分类

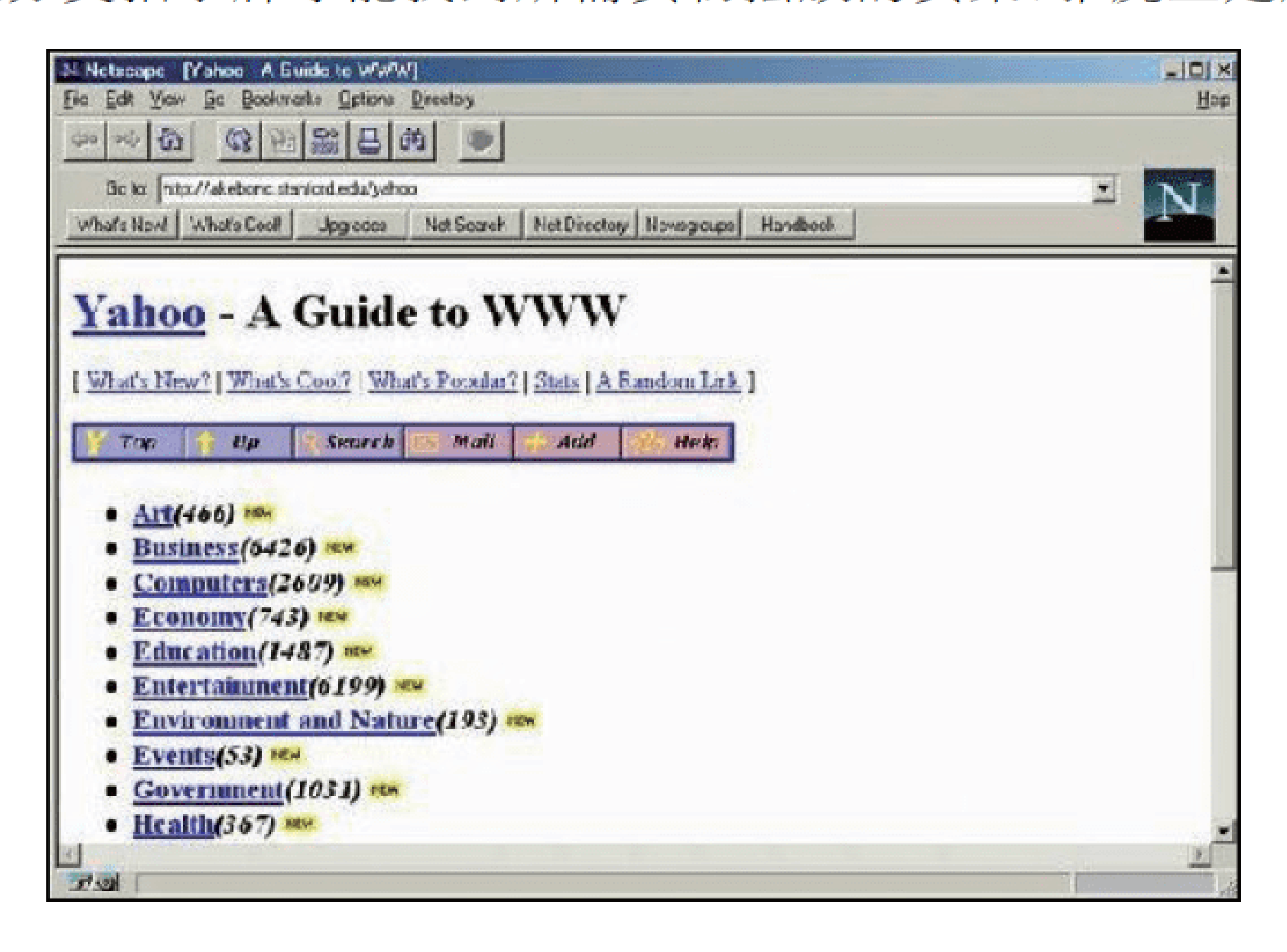


图 5.10　Yahoo 1994 年的界面

指示牌的作用，为我们寻找“货物”指明了方向。

随着互联网上的信息爆炸式增长，分类目录越来越力不从心，从图 5.10 中可以看到，Business 类别中有 6426 个不同的网址，用户要从这么多不同的网址中寻找自己想看的内容实在是太困难了。即使已经收录了这么多网址，仍然有绝大多数的网址不能被收录。另一方面，分类目录只能告诉用户信息存放在哪里，却难以显示出信息的细节。如果你想在沃尔玛买几斤澳洲大龙虾，通过货架上方分类指示牌只能知道虾摆放在哪个货架上，至于那里到底有没有大龙虾，大龙虾是否是澳洲产的更是无从得知。你必须穿过长长的过道，走过一排排货架，来到摆放虾的货架，经过一番肉眼搜寻后，才能知道到底有没有你想要的澳洲大龙虾。这是多么糟糕的体验啊！

面对分类目录的缺陷，用户渴望一种更好的搜索工具。用户渴望搜索工具能够像哆啦 A 梦那样神通广大，全世界所有的东西都装在它的兜里，只要说要一个名字，哆啦 A 梦立刻从兜里掏出来，而不用像在超市货架上找货物那样费时费力，运气不好时还极有可能找不到，无功而返。

1995 年 3 月，搜索领域的两位传奇人物谢尔盖 · 布林（Sergey Brin）和拉里 · 佩奇（Larry Page）在斯坦福大学相遇了。对，没错，又是斯坦福大学！当时正值互联网大潮，两人决定从信息检索技术入手打造一番事业。布林和佩奇想创造这样一个工具，它能收录全世界所有的网页，只要用户发出查询词，能马上从海量的网页中找到用户需要的那个。

在布林和佩奇之前，有人将文本检索用到了互联网搜索领域，AltaVista、Excite 就是这类搜索引擎的代表。他们采用了经典的信息检索模型，例如布尔模型、向量空间模型或者概率模型，来计算用户查询关键词与网页文本内容的相关程度。与分类目录相比，这种方式能收录大部分网页，并能按照网页内容和用户查询的匹配程度进行排序。然而，文本检索有一个很大的问题，就是搜索质量不是很好。

为什么文本检索质量不好呢？我们来分析一下其中缘由。文本检索只考虑网页内容和用户查询关键词的匹配程度，却不关心网页内容的质量。通过下面的例子你就会明白这是一件多么糟糕的事情。有一位未婚姑娘请一个婚介公司为她介绍未婚男子，姑娘提出的要求是又高又帅，那么如果婚介公司把身材最好、长相最帅的男子推荐给她就是最好的推荐吗？呵呵，这可不一定哦！因为除了身高和长相，对一个男子的评价标准还有很多，包括年龄、学历、家庭背景、职业、性格、收入等，也许身材最好、长相最帅的男子是一个又穷又笨又懒的男人，那么姑娘还是会对这个推荐非常不满。同样道理，与用户查询词最匹配的网页未必质量最佳，用户仍然有可能对这个文本上所匹配的查询结果感到失望。基于文本检索的搜索引擎显然不是大雄想要的哆啦A梦，试想一下，如果大雄想吃苹果，哆啦A梦从兜里掏出了一只腐烂发臭的苹果，大雄会是什么心情！

如何才能找到文本内容既匹配，质量又很靠谱的网页呢？布林和佩奇觉察到文本检索方式遗漏了一个重要的信息。互联网上的海量网页并不是孤立的，它们之间通过链接彼此相连，这些链接关系一定隐含着某种意义，而文本检索从未利用过这些链接关系。也就是说，信息不仅存在于网页之中，互联网中网页之间的逻辑关系也隐藏着很多对我们有用的信息。

佩奇想出了一个名为 PageRank 的奇妙算法。这个算法后来成为人类抗击信息泛滥的一件强有力的法宝，它使搜索引擎的质量大幅提升，堪称搜索引擎算法的开山鼻祖。1999 年佩奇发表的关于 PageRank 算法的论文迄今已被引用超一万次。汤森路透（Thomson Reuters）公司在 2014 年 10 月所做的一项统计表明，人类历史上发表的总共 5800 万篇 SCI 检索论文中，仅有 14 499 篇论文的引用次数超过 1000 次，而引用次数达到 12 119 次便可跻身百强榜。PageRank 算法在学术领域的重要地位可见一斑。

那么，这个传奇的 PageRank 算法到底是怎么工作的？我们先从一个人

物影响力排名的案例来领悟 PageRank 的核心思想。如图 5.11(a)所示，我们要对 A、B、C、D、E 等 5 个人的影响力排名。我们用箭头表示这 5 个人之间的认识关系，例如 A 有两个箭头分别指向 B 和 D，这表明 A 认识 B 和 D。注意这里的认识关系是单向的，例如我们大家都认识美国总统特朗普，可是他并不认识你。

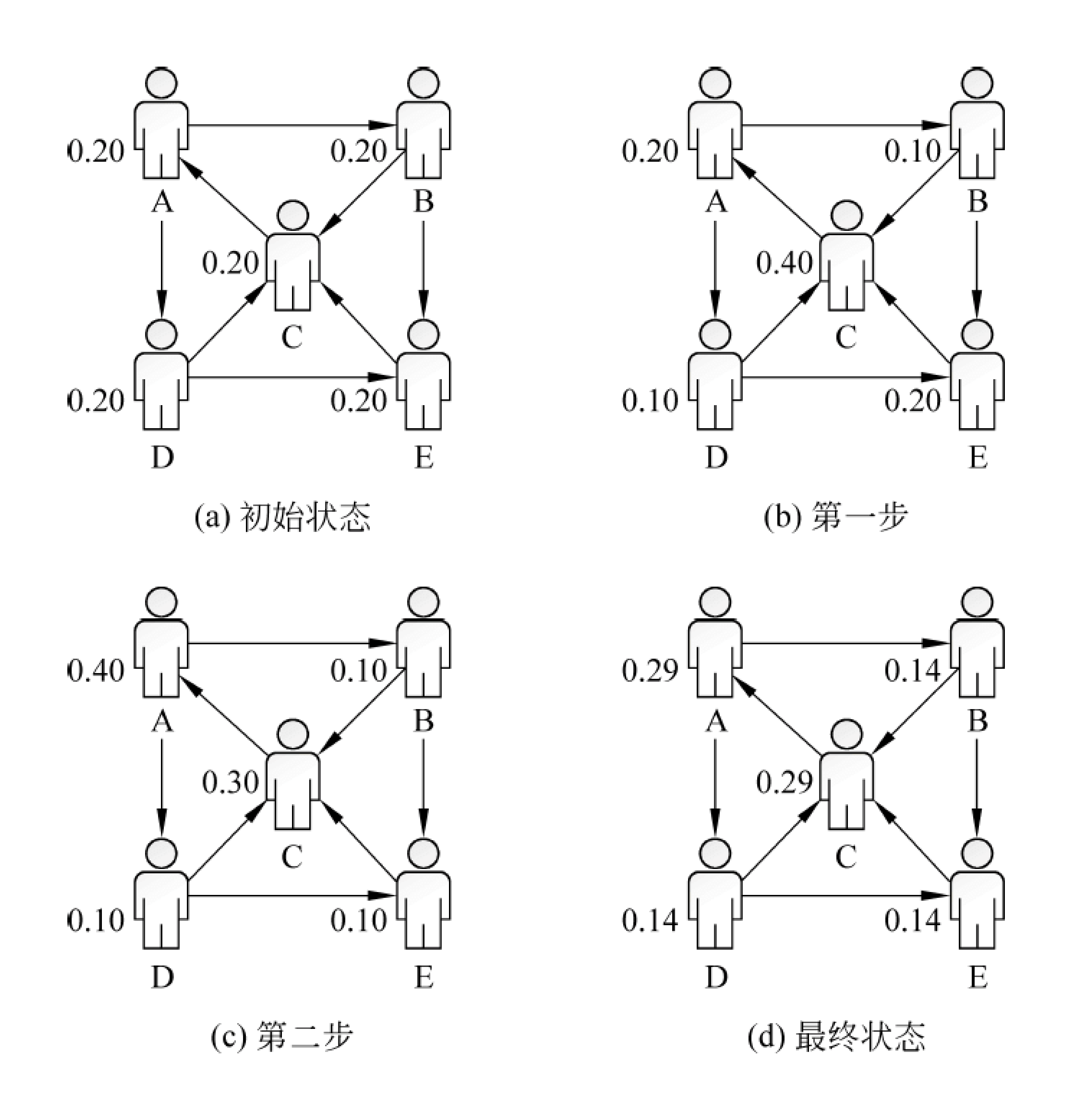


图 5.11　人物影响力排名

那么我们如何知道一个人的影响力到底有多大呢？

我们假定 5 个人的影响力之和是 1，每人的影响力值最初都是 0.2。我们刚才说过，被越多的人认识影响力就会越高，那么我们就制定这样一个规则，每个人都把自己的影响力值均分给自己认识的人。例如 A 的影响力初始值是 0.2，A 认识 B 和 D，那么 A 就把自己的影响力值均分给 B 和 D，各给 0.1。

按照这样的算法，我们就能对 5 人的影响力做排名了。具体计算过程是这样的，如图 5.11(a)所示，5 人的影响力初始值都是 0.2；在图 5.11(b)中，每个人都把自己的影响力值均分给了所有自己认识的人，于是 A、B、C、D 和 E 的影响力值分别更新为 0.2、0.4、0.1、0.1 和 0.2；我们再把图 5.11(b)中每个人的影响力值作为初始值，重复上述影响力值的均分过程，就会得到图 5.11(c)所示的结果；不断重复，5 个人的影响力值最终不再变化，分别是 0.29、0.29、0.14、0.14 和 0.14。这时，我们发现 A 和 C 的影响力明显高于其余三个人。

通过上面的计算，我们可以得出两个规律：第一，认识你的人越多，你的影响力就越大。在上述 5 个人中，认识 C 的人高达 3 个，而其他人仅被 1～2 个认识，难怪 C 的影响力会这么高；第二，认识你的人影响力越大，那么你的影响力也越大。尽管 A、B、D 都是只被一个人认识，但 A 的影响力却比 B、D 都高，甚至连被两人认识的 E 都比不过 A。

上面的案例利用人物之间的认识关系来计算人物影响力，而刚刚提到的 PageRank 算法则是利用网页之间的链接关系来计算网页权威值。如果把图 5.11 中的 5 个人替换成 5 个网页，把认识关系换成链接关系，那么网页 A 就有两个链接分别指向网页 B 和 D。网页之间的链接关系和各自的权威值如图 5.11(a)所示，多次重复人物影响力排名中的迭代过程，5 个网页的权威值就会收敛到图 5.11(d)所示的结果。

你是不是觉得这个算法计算过程很简单，简直不敢相信它就是成就了 Google 搜索霸业的神奇算法！图 5.12 是 Google 在 1998 年推出的 Beta 测试版本的界面，与 Yahoo 对比一下，目录进化成了一个输入框，你想查什么就在输入框里输入什么，剩下的就交给 PageRank 算法了。终于我们第一次进入了可以自由搜索的 Web 世界。

到目前为止，学术界已经提出了很多种基于链接关系分析的搜索算法，



图 5.12　Google 1998 年的界面

除了 PageRank 外，另一个具有代表性的链接分析算法称为 HITS。

HITS 算法的全称是 Hyperlink-Induced Topic Search，HITS 算法是由康奈尔大学的 Jon Kleinberg 博士于 1997 年首先提出的。HITS 几乎与 PageRank 同一时期被提出，HITS 算法同样以更精确的搜索为目的，直到今天仍然是一个优秀的算法。HITS 算法到底是如何工作的？为了便于理解，我们还是暂且不谈网页搜索这么深奥的问题，而是从一个故事中领悟 HITS 算法的精妙所在。

话说北京城里有四位爷们打算一起在外面撮一顿，可是最发愁的就是去哪儿吃。他们决定从东来顺、大鸭梨、黄记煌、必胜客和海底捞中选一家，可是这 4 个人意见并不一致。最后只能投票来决定，哪家得票最高就去哪家。图 5.13 就是他们的投票情况，有 3 人选东来顺，3 人选大鸭梨，1 人选黄记煌，2 人选必胜客，2 人选海底捞，所以这 5 家餐馆的得票分别是 3、3、1、2、2，也就是说东来顺和大鸭梨都是 3 票，并列第一，还是没法确定去哪一家。

这时，二爷有个提议，他说："咱们每个人品位高低不一，不应该把 4 个人的选择都看成一样的，要重点考虑品位高的那个人的推荐。"因为这 4 个人都自诩是吃遍天下无敌手的资深吃货，在品位上谁也不肯服谁。当大爷、二

	大爷	二爷	三爷	四爷	餐馆得分	
东来顺	♡		♡	♡	3	21
大鸭梨	♡	♡	♡		3	20
黄记煌		♡			1	6
必胜客	♡			♡	2	15
海底捞		♡		♡	2	13
推荐者“水平”	8	6	6	7		

图 5.13　餐馆选择问题

爷、三爷正吵得不可开交的时候，四爷突然有了一个主意：“一个人推荐的餐馆越受欢迎，说明他的水平越高，对不对？”大家点头表示赞同。四爷接着说：“好，既然大家都同意，那我们就用一个人所推荐的餐馆的得票总数来评价他的水平，怎么样？比如二哥，共推荐了 3 家餐馆：大鸭梨、黄记煌和海底捞，这三家餐馆的总得票是 3＋1＋2＝6，那么他的水平就是 6 分，同样道理，大哥、三弟和我的水平分别是 8 分、6 分和 7 分。”用所推荐的餐馆的得票总数来评价推荐人的水平确实是个客观公正的好办法，大家都对这个方法心服口服。于是大家开始重新计算各餐馆的得分，这一次将推荐人的水平作为权重来计算每家餐馆的得分。比如东来顺的得分是 8＋6＋7＝21，同样道理，大鸭梨、黄记煌、必胜客和海底捞的得分分别是 20、6、15 和 13。这样东来顺就比大鸭梨领先 1 分，最后他们就高高兴兴地去东来顺吃涮羊肉啦！

这个选餐馆的故事，通俗易懂地讲清楚了 HITS 算法的精妙之处。那么 HITS 搜索算法实际是如何运行的呢？

假设我们要搜索的关键词是“newspaper”。图 5.14 中左边是与“newspaper”字面相关的网页，右边是它们所指向的网页，得到的“票数”是指共被多少个网页所指向。如同在餐馆推荐中可以用餐馆的得票数反过来评价推荐者的水平，如图 5.15 所示，我们也可以用网页的“票数”反过来评估左边指向它们的“推荐者”的分量。然后，我们就可以像选餐馆时将推荐水平作为权重来计算餐馆得分那样，如图 5.16 所示，也考虑推荐者分量，用加权评

分的方法重新评估右侧网页的得分。我们用图 5.16 中右侧网页的新得分再次反过来评价左侧“推荐者”的分量，然后用新的“推荐者”分量加权评分会得到右侧网页的新得分，如此循环反复下去，右侧网页的得分就会收敛到某个固定值。最终搜索引擎就会把得分最高的几个网页呈现给用户。

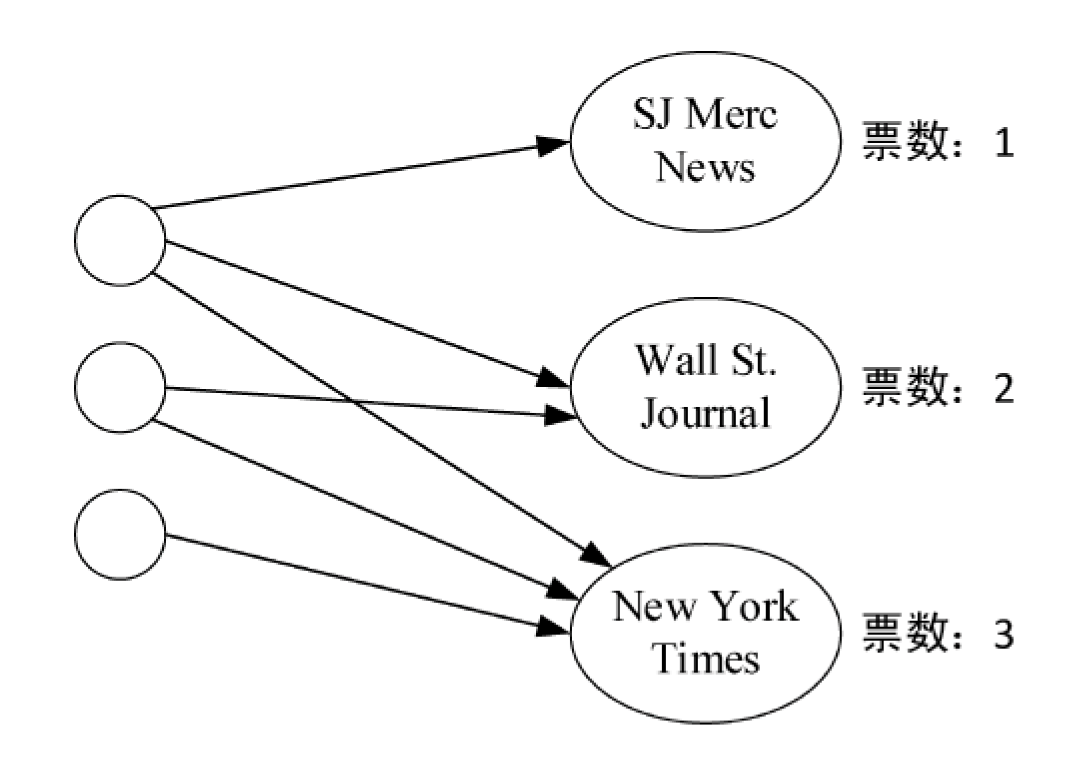


图 5.14　HITS 算法的第 1 步

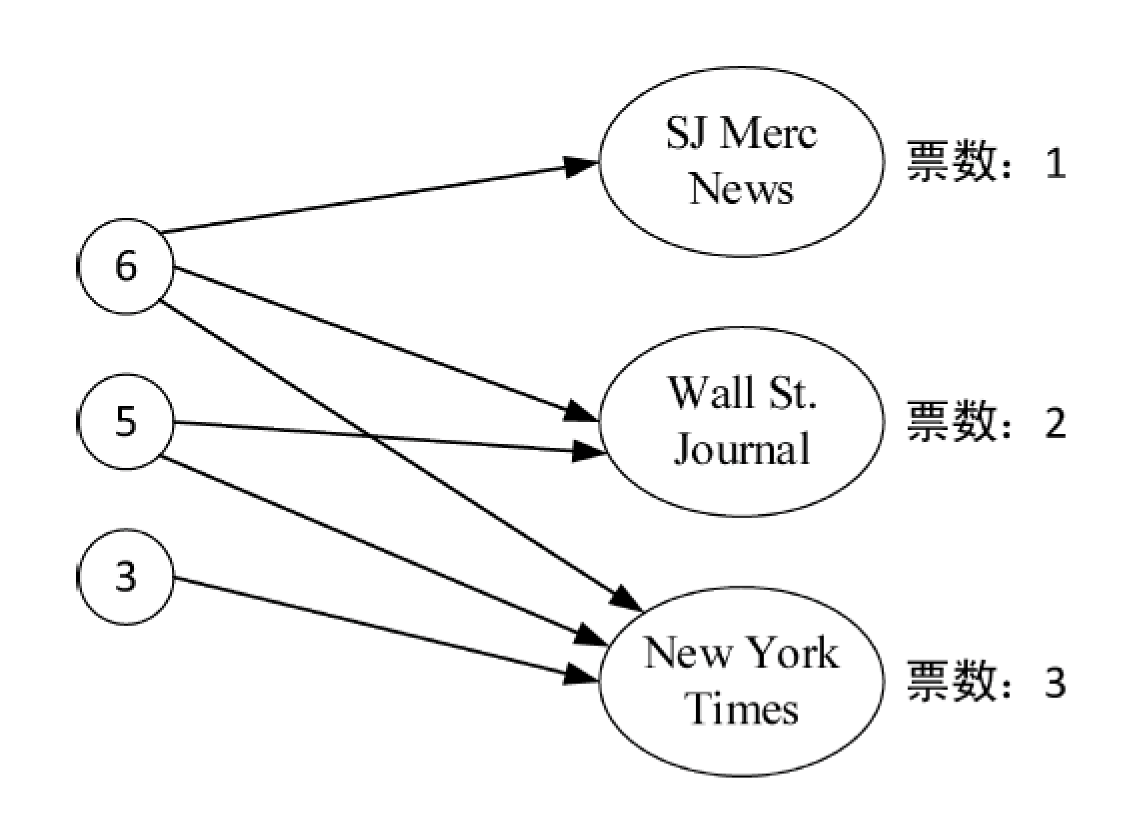


图 5.15　HITS 算法的第 2 步

这就是 HITS 算法的全过程。我们发现它与餐馆推荐方法唯一的区别就是要循环反复几次罢了。

HITS 算法认为每个网页都具有两面性：权威性和中枢性。如果一个网页被很多网页指向，表明其权威性高，认可度高；如果一个网页指向很多网页，表明其中枢性强；如果一个网页被很多中枢性强的网页指向，权威性更

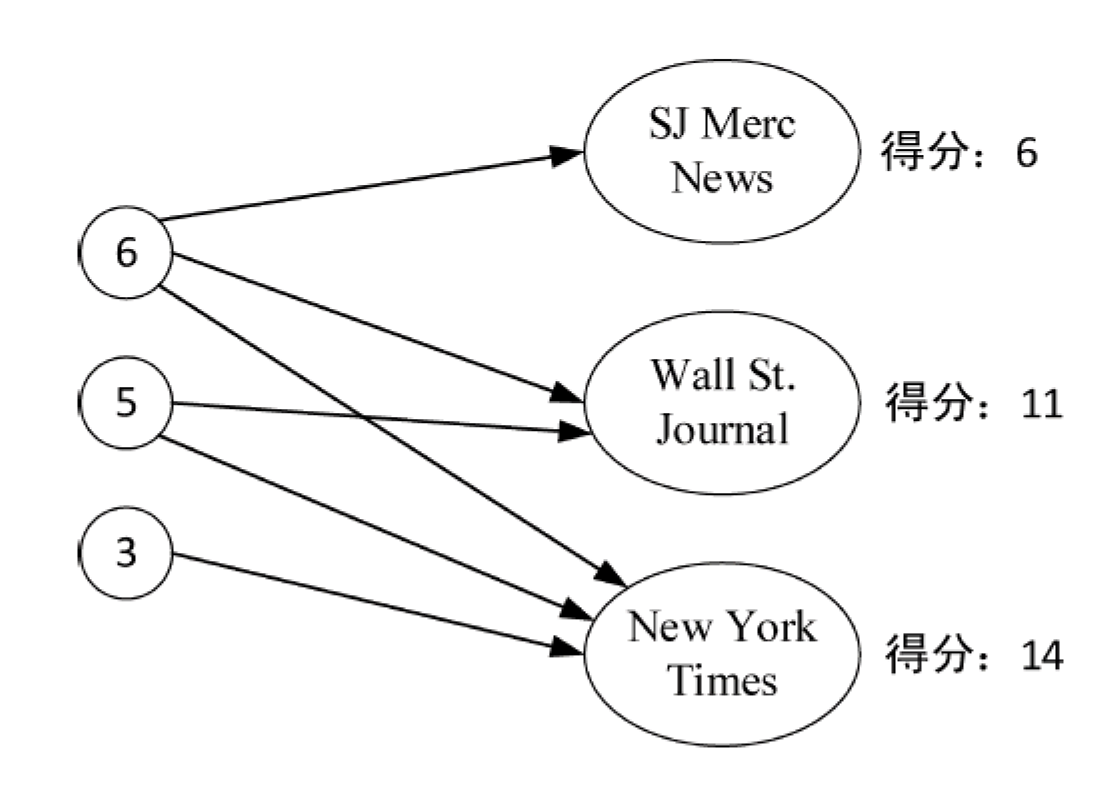


图 5.16　HITS 算法的第 3 步

高；如果一个网页指向很多权威性强的网页，中枢性更强。权威值其实就是图 5.14～图 5.16 中右侧网页的得分，中枢值其实就是左侧“推荐者”的分量。HITS 算法最精妙之处就是用权威值反过来计算中枢值，再用中枢值加权计算权威值，循环反复，直至收敛。

作为搜索引擎链接分析的两个最重要的算法，HITS 算法和 PageRank 算法无论是基本概念模型，还是计算思路及技术实现细节都有很大不同。那么 HITS 算法和 PageRank 算法各自有何优势和劣势呢？首先，HITS 算法是与用户输入的查询请求密切相关的，而 PageRank 与查询请求无关。所以 HITS 算法可以单独作为相似性计算评价标准，而 PageRank 必须结合网页内容的相似性计算才能对网页相关性进行评价。其次，HITS 算法因为与用户查询密切相关，所以必须在接收到用户查询后进行实时计算，计算效率较低；而 PageRank 则可以在爬虫抓取网页内容后离线计算，运行时可以直接使用计算结果，计算效率较高。第三，HITS 算法的计算对象数量较少，只需计算扩展集合内网页之间的链接关系；而 PageRank 是全局性算法，需要对所有互联网页面节点进行处理。因此，从两者的计算效率和处理对象集合大小来比较，PageRank 更适合部署在服务器端，而 HITS 算法更适合部署在客户端。

推荐算法——给你想要的一切

本章一开始提到了互联网推荐系统，它可以根据互联网用户的各种信息，向用户推荐未来可能感兴趣的信息。互联网推荐系统的灵魂，就是推荐算法。

对大部分人来说，因为音乐的存在，那些枯燥的时光开始变得轻松而美妙。现在我们每天只需通过一个网络音乐软件，就可以收听到海量的歌曲，有时候这些歌曲似乎明白你的心思一样，播放出的每一首歌曲都让你怦然心动。这时你也许会通过音乐软件的界面点一下“评价”按钮，表示喜欢，当然，有时候你也会选择“删除”，表示不喜欢。

现在我们有很多这种个性化音乐电台，它们通过基于内容的推荐算法，利用乐曲之间的相似性来为用户做推荐。这个算法的原理很简单，但是它最大的挑战在于如何计算歌曲的相似度。一方面，歌曲的数量庞大，而且每天还有大量的新歌诞生，分析这些歌曲的特征是一个浩大的工程，需要高效的大数据处理算法支持。另一方面，音乐的特征包含诸多方面，比如旋律、节奏、编曲、歌词、风格等，人类对音乐的体验是一个非常复杂的过程，很难使用自动化的程序来帮人类试听和分辨乐曲的特征并做标记，因此主要依靠人工的方式。

2000 年 1 月 6 日，著名的音乐电台 Pandora 曾经开展了一个浩大的音乐基因工程项目。音乐家和研究人员亲自听了上万首来自不同歌手的歌，然后对歌曲的不同特性（比如旋律、节奏、编曲和歌词等）进行标注，这些标注被称为音乐的基因。然后，Pandora 会根据音乐基因来计算歌曲的相似度，并向用户推荐与其之前喜欢的音乐在基因上相似的其他音乐。

下面我们再来看看基于人口统计学的推荐算法。当我们在网站上注册

时，经常会需要填写手机号码、电子邮箱、性别、年龄、职业等个人信息。我们很少去想网站收集这些信息会有什么用，我们只想迅速填写完，尽快完成注册。我们不曾想到，就在我们提交这些信息的一刹那，网站的老板在暗处偷偷地笑了，基于这些信息，推荐算法已经知道怎么把东西推销给我们了。从另外一个角度来看，这些信息实际上暴露了我们的个人隐私，可能会给我们带来不必要的麻烦。关于隐私保护，本书第7章会详细介绍。

仅仅利用这点简单的信息就可以吗？也许你对此并不确信，将信将疑，你可能认为，即便有点用，其效果也不会太好。其实，这些信息远比我们想象的有价值，性别、年龄与兴趣的相关性也非常高。女生们爱看有长腿欧巴的《来自星星的你》，而男生们更爱看惊险刺激的《越狱》。尽管从外表不能明显地区分出80后和90后，然而一旦80后和90后一起去K歌，年龄就暴露无遗。90后唱的歌，80后压根没听说过，即便容颜未老，80后也已然感觉年代的鸿沟！

利用人口统计学特征包括年龄、性别、工作、学历、居住地、国籍、民族等做推荐的算法，称为基于人口统计学的推荐算法。这一算法是利用用户之间的相似度来做推荐的。如果小兰和大兰都是年龄25～30岁的年轻女性，那么她们的兴趣有可能相似，把小兰喜欢的电视剧推荐给大兰，那么大兰有可能也会喜欢。

基于人口统计学特征的推荐系统的典型代表是布鲁斯·克鲁维斯(Bruce Krulwich)开发的Lifestyle Finder系统。首先，克鲁维斯将美国人群根据人口统计学属性分成62类，每个注册用户都需要填写个人资料。算法根据用户资料判断他属于哪个分类，最后给他推荐这类用户最喜欢的15个链接，其中5个链接是推荐他购买的商品，5个链接是推荐他旅游的地点，剩下的5个链接是推荐他去逛的商店。

为了证明利用用户人口统计学特征后的推荐结果好于随机推荐的结果，

克鲁维斯做过专门测试。相对于利用人口统计学特征的算法，他设计了一个实验组和一个对照组，实验组看到的推荐结果是个性化推荐算法生成的，对照组用户看到的推荐结果是完全随机的。实验结果显示，对于人口统计学特征的个性化推荐算法生成的推荐内容，用户点击率为 89%，而随机算法推荐内容的点击率只有 27%。对于利用人口统计学特征的个性化算法，44%的用户觉得推荐结果是他们喜欢的；而对于随机算法只有 31%的用户觉得推荐结果是自己喜欢的。因此，我们得到一个结论：使用人口统计学信息相对于随机推荐能够获得更好的推荐效果。

基于人口统计学的算法有一个重要的用途，就是解决系统的冷启动问题。对于一个新注册的用户，系统无法根据他的历史行为，如购物车记录、购买记录、评价留言推测他的兴趣，此时利用新用户注册时填写的性别、年龄、职业等信息就可以做最初的推荐，效果比漫无目的的随机推荐强太多了。随着用户使用时间的增长和历史数据的积累，其他的推荐算法才能有用武之地，从而继续提高推荐的精度。

与基于人口统计学特征的推荐算法类似，微信精准广告推送也是基于对用户特征的深刻了解。据微信称，从推送对象的角度，其广告引擎从“高活跃度”“常参与广告互动”两个评分维度中精选了一批“种子用户”作为广告的第一批推送对象。以他们为中心，挖掘出一批和他们兴趣相同的好友。这种推送策略大大减少了广告对用户体验地消极影响。从传播逻辑的角度思考，你的好友对广告的互动情况将影响你接收同一广告的概率。例如，你和小马哥是好友，小马哥看到某条广告，点赞或评论了它，那么你看到这条广告的概率就会提升。从推送策略上来看，一条朋友圈广告在曝光后 6 小时内若未产生互动，那么这条广告将自动消失；如果曝光后有互动，那么广告不消失。一个广告将持续 7 天有效，对单个用户每 48 小时内只推送一个广告。以上几条政策从不同角度来保证用户体验，同时提高了受众定向的精准度。

大数据算法的新征程

计算机天生就是用来解决以前没有遇到的问题的。

——比尔·盖茨，微软公司创始人

滞后的大数据处理能力

与农业时代、工业时代甚至互联网时代初期的数据相比，赛博时代的数据规模、复杂程度、产生速度、蕴含的信息量均有质的差别。一方面，移动通信技术、物联网技术、IPv6等新技术从根本上改变了数据产生方式，数据以前所未有的速度和规模产生，处理大数据所面临的难度更大；另一方面，每个人都是数据的生产者和消费者，人类的经济活动、科学研究、生产生活和数据的关系比以往任何时候都更加密切，对大数据处理提出了更高的要求。然而，目前的数据处理能力还不能完全满足需求，尤其到了赛博时代，矛盾更加突出。

与几十年前相比，经济活动参与者的数量激增，我国上市公司数和上市股票数都保持了高速增长。根据2017年1月24日的数据，沪市有上市公司1205家，仅A股就有股票1198支，市价总值291 743.54亿元，每天的成交量就有1 384 397万股，成交金额1408.41亿元，共计成交664.27万笔。也就是说，沪市开市时，每秒钟都有4.6万笔交易成交，成交金额约1000万元。股市的大盘波动信息、股票交易信息、股票价格信息等构成了一个超大规模的数据集合，再加上参与者行为数据的复杂变化，令市场的走势变化更加难

以把握，对算法的要求也就更高。想要处理这些规模更大、复杂度更高的大数据，算法必须具备快速采集数据、快速处理数据、快速分析数据的能力，并在此基础上能够快速做出正确决策。

以国家的长期宏观经济政策为例，政府通常需要评估当前的经济状况。在理想状况下，我们希望经济系统可以实时统计全国正在发生的所有经济活动，根据这些经济活动快速生成各项经济指标，这样才能真正准确地反映出当前我国的经济状况。当然，就目前的技术手段来说，这肯定是不现实的。当前常用的评估经济状况方法有很多种，国民生产总值（Gross Domestic Product，GDP）、消费者价格指数（Consumer Price Index，CPI）等都是我们耳熟能详的经济状况评估指标。那么这些指标到底能不能准确反映出国家当前的经济状况呢？答案是否定的。

以 GDP 为例。GDP 是一个国家所有常住单位在一定时期内生产活动的最终成果。GDP 的计算过程高度复杂，从覆盖面来说，GDP 数值受到国民经济各行业经济活动的影响，涵盖政府投资、金融交易、进出口、货物买卖等诸多方面；从数据来源来说，既需要统计部门提供的数据，还需要财税、金融、保险等各相关部门的数据；从计算方法来说，GDP 数值并不是简单相加或者汇总得到的，而是通过经济普查、抽样调查、虚拟计算等多种手段综合得出的。在这种情况下，国家通常每季度发布一次 GDP 初步核算结果，发布时间通常滞后两个月左右。每年度对经济数据核实后，或者开展全国性经济普查活动后，还有可能根据情况对已公布的 GDP 结果进行修订。类似的还有每月公布的 CPI 指数，尽管 CPI 指数在公布后通常不会被修订，但是其公布时间通常滞后三周。

在宏观经济层面，我们尚缺乏快速有效的数据采集、处理、分析手段，还需要大数据处理技术的进步和发展。实时性和准确性就是其中最为常见的冲突需求。由于数据规模过大、数据复杂程度过高，对数据进行完全精确的

统计和计算需要的时间可能是一个天文数字；等到我们得出了精确的计算结果，这个结果的时效性已经很差，也就没有很高的存在价值了。所以，通常为了保证在可接受的时间内计算出结果，就要适当放松对数据准确性的要求，这样做的结果就是，我们得出的结果往往既是滞后的，又是不准确的。我们就如同还没有结绳记事技术的原始先民，在等待这个时代的神农氏。

算法竞赛的风险

虽然我们的大数据处理能力还不能与处理需求相称，但是不可否认的是人类从未因此停止努力。

在电影中我们看到有关股市的场景总是充满了紧张与刺激，计算机屏幕上显示着快速跳动的股票价格，股市大鳄目不转睛地盯着不断变化的股票指数，不停地用红蓝笔和尺子在纸上验算着价格走向，急促地打着交易指示电话，甚至来不及咬上一口放在桌上的汉堡，因为很担心一个汉堡的功夫就会让户头上的数字产生心惊肉跳的变化。

当然，电影对现实生活有升华和夸张的成分。在现实中的股市里，大鳄们如果真的是这样炒股，先不说身体会累垮，在连续高强度的工作下，人怎么能保证做出客观、精准的决策呢？况且资本市场瞬息万变，操作人员打一个哈欠或上个厕所都可能浪费宝贵的时机；大鳄发现机会后还要用电话通知手下下单，这样层层传达所花费的时间也很可能会错失赚钱良机。金融家们怎么能容忍这样的事情发生呢？

实际上，股市大鳄们的雇员们根本不是精通经济学和金融投资的专家，在股市的另一端，与股市散户们对弈的是一台台被算法赋予神奇能量的计算机。自20世纪70年代起，就有人探索通过计算机实现全自动的股票交易，也就是所谓的程序化交易技术。随着计算机技术的发展，股票交易所中喧嚣

的交易人群、忙碌的黄马甲更是早已不见踪影，甚至从某种意义上来说，人类已经成为华尔街的看客。在跳动的指数和蜿蜒的曲线背后，是一台台计算机和一行行的算法代码在发挥作用。这些计算机按照计算机专家们事先设定的算法，根据资本市场的走势快速做出判断和决策，发起动辄数千万美元的交易。1986 年率先采用程序化交易的托马斯·彼得菲（Thomas Peterffy）目前已经是全美排名前 100 的大富豪。彼得菲也是 Interactive Broker（盈透，简称 IB）的创始人。作为美国最大的非银行券商，从 1977 年成立以来，IB 就一直全心投入于程序交易科技的研发，目前 IB 每天处理超过全世界 14％的股票期权交易量。

如今，美国交易市场有超过 60％的交易都是计算机程序完成的。计算机的阅读速度、分析速度和操作速度远超人类，最重要的是计算机是绝对理智的，永远不会被数字冲昏头脑，永远不会因为精神疲劳做出错误决策。股市中隶属于不同公司的计算机按照各自的算法紧张运转，在资本市场这个虚拟的战场上激烈交锋，带动各种经济指数波动前行。

当越来越多的计算机算法参与到经济活动中来之后，计算机的决策方式、计算速度、记忆力、理性程度等都和人类截然不同，市场走势的变化规律也随之发生变化。与前面介绍的谷歌公司推出的流感趋势预测类似，在情绪的驱使下人们行为模式会发生变化，而算法所了解的局面总是表面的、局部的，因此算法做出的决策也并不会永远正确。雪上加霜的是，当更多的公司借助计算机算法参与到自动交易中后，为了战胜其他公司的计算机算法实现盈利，算法之间的攻击和防守愈演愈烈，“瞒天过海”“暗度陈仓”的手段层出不穷。计算机算法的功能越来越复杂，程序规模也越来越庞大，随之而来的是算法可靠性的降低。程序的错误会给市场走势带来严重影响，并且影响的恶劣程度远远超过任何利空消息带来的影响。

2012 年 7 月底，美国著名的做市商骑士资本集团（Knight Capital

Group)对其与纽约证券交易所相连的程序化交易系统进行升级,安装了一个新的软件模块。然而,负责软件升级的员工忘记对一台服务器上的废弃软件模块进行更新。这个小小的疏忽,竟然给骑士集团带来了始料未及的严重后果。2012 年 8 月 1 日,纽约证券交易所开市后的 45 分钟内,骑士集团的程序化交易系统发起了几百万笔交易,其中有 400 多万笔成交,成交的 4 亿股股票的交易金额高达 66.5 亿美元。这些由于程序错误发起的异常交易导致多支股票价格发生离奇波动,甚至触发了涨跌阈值,停止交易。此次失误,引发了投资者对于骑士资本集团的强烈质疑,并最终导致该集团被另一家做市商 Getco 吞并。

2010 年 5 月 6 日,美股发生的"闪电崩盘"事件也印证了计算机算法的脆弱性。这一天,美国市场并没有什么特别的利好或利空消息,各项经济数据也没有什么值得太多关注的地方。然而,从 14 点 42 分开始,美股如坐了一次惊心动魄的过山车。短短 5 分钟时间内,道琼斯指数从 10 458 点瞬间跌至 9869.62 点。到 14 点 58 分,道琼斯指数又回到 10 479.74 点。这次崩盘造成了极为恶劣的影响,埃森哲(Accenture)咨询公司开盘报价为 41.94 美元,14 点 50 分时股价竟然跌至零美元,后收于 41.09 美元。这一天,与埃森哲有类似经历的上市公司不在少数。这次崩盘给了华尔街当头一棒。美国证券机构专门对此事进行了紧急调查。然而,调查并没有确认问题的源头,疑似的原因之一,是花旗的一位交易员在执行交易时,误将"M"(million,英文意为"百万")错输为了"B"(billion,英文意为"十亿"),这个事件触发宝洁的某个道指成分股急剧下跌,进而引发某些公司的计算机算法发起了自动交易。而不同公司的计算机算法本就处于竞争关系,算法之间又相互触发、相互影响,最终竟形成雪崩式的暴跌局面。对此,人们普遍认为参与程序化交易的计算机算法是加剧崩盘的元凶。

也许现在就把我们的世界完全交给计算机算法还为时尚早。然而,这并

不意味着我们应该抛弃计算机算法，反而更加凸显了大数据处理算法在赛博时代的重要性。我们要做的是不断开发更为强大和更为可靠的算法。从当前的发展现状来看，大数据处理算法尚未成熟，还有很多没有解决的问题，搜索、推荐、社交网络等各方面都还有进一步改进的空间。或许深度学习和人工智能技术的引入会使大数据算法获得第二次腾飞。

作为智能经济时代的第一生产力，算法无疑会帮助我们更加深入地审视经济系统整体的运行状况，进而制定更加及时有效的经济政策。也许我们不应该将大数据仅仅看做一个单纯的技术问题或者社会问题，因为它正在真切地影响着我们如何观察和思考这个世界，在这个人与数据共存的时代中，或许我们应该把大数据当作一种新的世界观，毕竟，无论人类怎么想，大数据就在那里。

第 6 章　你的数据究竟该卖多少钱

世界上最大的零售商,没有自己的产品库存;世界上最大的内容提供商,没有自己出版的内容;世界上最大的租车公司,没有自己的出租车;世界上最大的出租公寓,也没有自己的房子。这些智能经济特有的全新模式得以实现,关键就在于数据。我们这个时代的数据正在对经济产生前所未有的深刻影响。就在本书编写即将完成的 2017 年 6 月,菜鸟与顺丰发生的数据接口之争充分说明了这一点。

菜鸟网络科技有限公司是 2013 年 5 月 28 日,由阿里巴巴集团、银泰集团联合复星集团、富春集团、顺丰集团、三通一达(申通、圆通、中通、韵达),以及相关金融机构共同合作组建的。作为中国最大的两家物流公司,菜鸟与顺丰之间的关系非常微妙。虽然顺丰和菜鸟有合作,但同时又在搭建自己的快递体系。2015 年 5 月,顺丰联手中通、申通、韵达和普洛斯一起投资 5 亿成立“丰巢科技”来与菜鸟展开直接竞争。

2017 年 6 月 1 日,菜鸟称,顺丰于当日凌晨关闭了对它的数据接口,并

停止给所有淘宝平台上的包裹回传物流信息。顺丰立即回应称，丰巢数据接口是菜鸟方关闭，因为顺丰拒绝了向菜鸟提供丰巢所有包裹信息（包括非淘系订单）的要求，并认为这是一次有针对性的封杀行动。顺丰方面认为“菜鸟之所以封杀顺丰，背后的原因是阿里方面一直希望顺丰从腾讯云切换至阿里云。”

也许菜鸟和顺丰“闹掰”是迟早的事。早在事发一年前，京东创始人刘强东就提到了两者之间的矛盾所在：“菜鸟物流就是在为几家物流公司搭建系统，说得好听是提升这几家快递公司效率，说得难听点，这几家公司的大部分利润都是被菜鸟物流给吸走。”而之所以能吸走快递公司的利润，本质在于菜鸟改变了快递公司的模式。中通公司曾在其风险提示中提到：“与菜鸟的合作可能会使其提升成本，削弱与终端客户的连接，甚至打破以往的商业模式。”这样下去，最后这些快递公司很可能彻底沦为“跑腿的”，这也是为什么顺丰作为菜鸟的早期发起人，最后却退出了菜鸟的原因。

菜鸟推出的第一个成功的产品是电子面单，统一了商家和物流企业的接口和标准，能实时计算出每个包裹的路由，经过哪个分拨中心、站点、收件员，同时直接反映在三段码上，并实时监控。菜鸟物流云本质是先让快递公司上阿里云，并做系统拆分、改造及整体架构修改，菜鸟可以基于日志进行全链路系统与业务监控，并输出其五级地址库，经过博弈、动态规划、协调资源来达成供应链全盘统筹，提供整体解决方案。而实现这些的必要前提就是掌控数据，数据的核心就是标准，“当物流公司所有信息和作业（包括包装标准），配合干线和仓配资源的标准，再加上阿里对于商流的掌控，菜鸟理论上可以形成垄断，拥有绝对话语权。”据 36 氪采访的业内人士称。

未来，快递公司的利润会不会被菜鸟吸走？当然是可能的。长远来看，谁控制数据，谁就掌握了定价权。例如“双十一”，阿里说快递定价多少就得多少，你如果不听，那你的一大半业务就没了。

当菜鸟打通整体物流数据时,不管是仓储还是运输的调度、运力分配,它都拥有话语权和定价权,所以你很难在这个行业拿到最大利润。“原本物流的主动权是供方的,现在已经完全转移到了需求方,之后会转移到数据上,从人找货变成货找人,物流、分销、零售的角色会因数据而改变”,银泰集团CEO陈晓东曾对菜鸟的数据应用如此评价。通俗地说,这就像套在脖子上的绳索,你的命掌握在别人手上,这样下去,未来快递公司之间的竞争将完全变成价格竞争。作为这些数据的创造者和服务的提供商,这样的结局对物流公司来说,看似不太公平。

当然,我国当前的物流行业仍然在发展中,数据之争也刚刚开始,对于那些想继续做强的物流公司,谁都不愿意轻易地将数据的控制权拱手相让。尽管这次事件由国家邮政局出面进行了调停,但是未来物流行业的竞争一定会越来越激烈,这场数据的战争也仍然将继续下去。行政干预当然是一种方法,但对于日益发展的新经济来说,这毕竟不是长久之计。

目前来看,为了保证数据提供方的安全和利益,保证市场的持续、有效运转,用数据交易实现数据共享是个多赢的选择。然而这样做的前提是需要让数据成为一种能够被市场调节的商品,让数据本身能够创造价值的同时,也具有价格。首先要面对的现实问题就是目前大数据应用缺少权威的交易中介,阿里想拿顺丰的数据匹配,顺丰想拿阿里的数据匹配,却没有一个公平、合理又安全的方法。随着数据的战略价值和经济价值越来越高,数据行业应该出现一个中立的第三方,能够通过规范化的市场机制来实现数据的流转和共享。

在数据变现这个领域,最成功的应用应该是流量变现了,这是目前大多数移动互联网公司让数据产生价值的做法。图6.1给出了互联网广告实现的“数据变现术”。

图6.1左边是流量变现,也就是“流量变成钱”的故事:将某网页中的广



图 6.1　数据变现

告位以 10 000 元的价格卖给某知名剃须刀公司。但是,这显然是个不明智的选择:该公司广告的受众大多数是男性;将广告无差别地展示给所有网页用户,是对流量的浪费。于是,我们想,如果能像图 6.1 右边那样,将男性受众留给该公司,收取 6000 元广告费,再将女性受众留下来,以 6000 元卖给销售化妆品的广告主,那么总收入将达到 12 000,比单纯的无差别流量变现要高。而与此同时,剃须刀公司和化妆品广告主也都很高兴,因为他们用了 60%的成本获得了同样的有效受众。

这个方法的思路其实很简单,但如果找不到能把男性用户和女性用户区分开的办法,这个思路也没法实际使用。那如何才能将男性用户和女性用户区别开来,又如何进一步知道他们有没有剃须刀和化妆品的需求呢?这就得依靠数据了。所以说,这多出来的 2000 元是数据资产挣的钱,我们称之为数据变现。

对于提供广告位的媒体网站来说,数据提高了广告位的利用率,为网站增加了收益;对广告商来说,数据提高了其投放精准度,使广告商看得见"经过"广告位的客户,有的放矢。数据的加工、碰撞、流通提升了数据的价值。

数据提供方、数据需求方、数据服务商等多方可以构建以数据开放、数据交易、数据分析为核心的综合性数据开放平台，从而打造一个生态圈。在传统行业，有技术资本化；在数据时代，也可以有数据资本化。据《2016年中国大数据产业白皮书》不完全统计，2015年我国大数据相关交易的市场规模为33.85亿元，预计到2016年国内大数据交易市场规模将达到62.12亿元，2020年将达到545亿元。数据交易需求大，政府也在支持并推进数据交易，各个地方交易所、交易中心相继成立，带动了整个数据交易市场。

数据定价是数据交易的核心技术难题。有关数据价值的更科学、更深层次的探究和挖掘，包括数据交易的商业运转模式的探索都才刚刚开始。

数据该如何买卖

数据科学家是懂得获取、清洗、探索、建模和解释数据的人，还要融合入侵技术、统计学和机器学习。数据科学家不仅要处理数据，还要把数据本身当作一个五星产品来对待。

——希拉里·梅森，Fast Forward Labs 创始人

萌芽中的数据交易产业链

2014年，“社工库”这三个字出现了，人们很容易望文生义，把它理解成一个公益组织。但其实，社工库是一个模拟搜索引擎页面构造的信息查询网站，分为“QQ密码查询”“QQ资料查询”和“开房记录查询”三个部分。其中，QQ密码和QQ资料的查询需要输入想要查询的QQ号，而开房记录查询则

需要输入相应的身份证号。当然，这样一个泄露隐私的网站很快就被人举报了。尽管被举报的社工库网站已经被封，但仍有多家社工库仍然暗中进行继续运作。各个社工库所涉及的信息资料也有些差别，除了 QQ 资料信息、身份证信息以外，有的社工库还会有 163 邮箱、126 邮箱等用户的信息。这些社工库只会免费开放一小部分信息查询内容，想要查询其他内容，则需要付费。此外，一些社工库网站还会贴出广告，表示可以出售自己掌握的整个数据库。其中一个社工库团队表示，自己全部展示出的数据库售价 1500 元，另有腾讯、阿里巴巴、中国联通、银行等各大网站和论坛的"影响极大"的数据库。其公布的数据库目录列表显示，多家大型网站的用户数据均被其掌握。该网站团队表示，列表中的数据"只是本人出售的数据其中的一小部分"。

技术推动了社会的进步，与此同时，也创造了新的犯罪手段。

2016 年 9 月 7 日，中央电视台主持人撒贝宁接到骗子电话，声称自己是北京市公安局的，撒贝宁涉嫌拐卖儿童。骗子虽然当场被识破，然而骗子是如何得到撒贝宁的联系方式的呢？新京报的记者曾暗访个人信息买卖，发现在网络上买到个人数据信息并非难事。例如，准大学生信息的价格为 0.3 元/条，其中包含姓名、性别、出生年月、入学时间、身份证号码、手机号码、家庭住址以及父母姓名和联系方式等。

其实，一个人从出生开始，他/她的各种信息数据就已经被泄露了。骗子利用这些数据搜索到可能的"被骗者"，然后广撒网。除了打电话行骗以外，骗子甚至会收集、分析出你的银行卡信息，假扮成服务员在餐厅里看你结账时输入的密码，然后以迅雷不及掩耳之势将你的卡在一个小黑盒里刷一下，你的卡就被成功复制了。暴露的密码让骗子可以随意取走你银行卡里的钱。

自 2000 年以来，随着中国通信业的快速发展，通过掌握个人数据而进行的电信诈骗就屡见不鲜。掌握了一个人的数据，就掌握了分析和了解一个人的手段。骗子的惯用伎俩不外乎是用他知道的数据骗取你的信任，引导你的

行为。如果我们不能管理好数据、善用数据，其危害也会随着数据的增长而变得后患无穷。

这里其实有两个问题，首先是用户应该能够通过技术手段保护自己的隐私数据，这一点我们将在第 7 章再继续讨论。其次是当用户不得不向网络运营商或者内容服务商提供一些隐私数据时，这些数据能够通过一种合法、合理且公平的方式来进行数据交易。实际上，互联网的全部商业模式都是基于用户隐私数据进行运作的，当我们免费使用搜索引擎带来的便利时，必须牢牢记住，世界上没有免费的午餐，搜索引擎需要我们的隐私数据来推送广告，从而挣得收入来维持它的运转。

下面我们先来看一看数据交易平台的发展现状。

2014 年 12 月 31 日，贵阳大数据交易所成立，随后，数据堂、Datamall 等数据交易平台相继出现，一时之间，“忽如一夜春风来，千树万树梨花开”。目前这些数据交易平台大概可分为三种类型：平台型、技术型和综合型。

平台型数据交易平台只负责提供应用编程接口进行数据交易，需求方按调用次数付费。有淘宝形式的，如数粮和数据宝，单纯地为数据接口提供商和需求方提供一个交易的平台；也有聚合应用编程接口形式的，如聚合数据、SHOWAPI、HaoService，这些平台会先将数据提供商的第三方接口进行技术统一，需求方抓取数据时，就只需要面对统一的接口，从而大大简化了需求方从不同数据提供商获取数据的复杂度。

技术型数据交易平台本身就是数据提供商，如数多多、发源地、大海洋，它们提供数据采集服务，数据以数据包形式出售。需求方可以向平台明确提出自己需要的数据内容，平台进行试采集，将采集的样品数据给需求方检验，如通过，则进行数据采集，将结果打包交给需求方。

综合型数据交易平台，顾名思义，业务类型丰富，数据可以以应用编程接口、数据包、定制、众包等不同形式出售，如数据堂、优易数据都是综合型数据

交易平台。

在这些平台中，数据主要以两种形式进行交易：应用编程接口和数据包。也就是说，用户可以通过购买应用编程数据接口的访问权限，以得到自主抓取数据的权利，而计费标准，有的按照抓取数据的条数，比如每抓取 1000 条数据多少钱，有的则按照接口开放时间计费，比如每个月多少钱，在这个月内可以无限抓取；用户也可以直接购买或定制完整的数据包，根据数据类型、数据包大小、需求情况等付费。

数据交易平台正在摸索中前行，效果也是初见其成。

2016 年 1 月 22 日，华中大数据交易所做成了一单截至当时国内最贵的一笔数据交易——数十万户企业信息，卖出了 26 万元的价格。这个数据库里有全国几十万家企业的司法、市场、社会等方面信息，原本这些数据分散在工商、税务等十几个部门，搜集起来非常麻烦。航天信息公司购买了这个数据库，获得了一个共享的权限，需要哪个企业的数据就直接检索，方便省事。当然，这里卖出的数据一定是经过脱敏处理的，去掉了用户隐私部分，这是最基本的要求。但是，尽管隐匿了隐私信息，仍然存在数据安全问题，比如数据泄露了，或者隐匿得不好，原数据被发现了，这就涉及数据隐私保护问题，我们将在第 7 章继续讨论。

“我们花 26 万元购买了几十万个企业信息的搜索权限，如需扩大范围，会继续追加投入”，航天信息是一家上市公司，国内电子发票龙头企业，业务覆盖 4 亿个人用户，企业用户接近 1000 万家。“我们拥有很多企业的经营数据，但是企业信息其他方面涉及不多”，华中大数据交易所搜集到的数据正好与航天信息互补。双方数据整合后，就更能对企业了若指掌。而这些数据的分析结果，在金融机构开展信贷业务，企业选择合作伙伴，相关部门制定企业政策时，都可以用来参考。

数据交易平台为需要数据的企业、组织或个人提供了数据，买方利用数

据达到自己的目的。这个过程好比是一家人要做一顿饭,数据便是基本食材,数据交易平台就是菜市场。有的菜市场售卖的是原始的刚从地里采摘出来的蔬菜,刚从水里捞出来的鱼,有的则是清洗干净的蔬菜和已处理的鱼肉半成品,剩下的事情便由买家自己来完成了。如果只是一顿家常便饭,那么可以自己动手做出一桌晚餐;但如果是一次家宴,家里又没有"大厨",再丰富的食材也做不出像样的大餐,那么就需要叫外卖,或者请一个会做饭的厨师来帮忙。

因此,除了需要能够提供"基本食材"的数据交易平台,我们还需要有一些外包服务提供商来提供数据处理、数据分析、数据管理和行业解决方案等数据相关服务,这便是数据服务商,例如 DataEye、TalkingData。它们有的用自己采集的数据,有的用需求方提供的数据,来实现数据服务。

数据服务商的业务通常可以分为数据管理平台、数据管家和行业解决方案。

数据管理平台用来挖掘数据中的商业价值、进行数据资产化,把分散的多方数据进行整合纳入统一的技术平台,对数据进行标准化和细分。用户可以将细分结果推向现有的互动营销环境。

数据管家从数据来源、建模、运营及后台处理等多个维度助力企业处理数据资产,帮助企业实现盘活企业数据资产、提升企业管理效率、增强数据变现能力、助推新商业模式等目标。

行业解决方案是提供产品战略、产品运营、数据运营和收益评估的综合性解决方案,帮助企业提高投入产出比,从而创造更多的商业价值。

有菜、有料、有人做,才能完成一道美味的佳肴。数据服务商补齐了数据商品的周边服务,细化了市场分工,让专业的人去做专业的事情,这样数据市场的良性产业链才能逐渐形成。

一个健康有序的数据交易市场需要哪些参与者

数据交易的需求日益增大，这些数据的需求方究竟利用数据来做什么呢？

最常见的数据需求便是估计一件事物的价值和对比多件事物的价值。对需求方来说，数据市场就相当于一个数据仓库。他们利用数据市场提供的数据，例如来自网页的“开放数据”，以及非公开源，例如来自路透社的商业数据和内部私人数据，来进行商业估值。需求方制定描述“事物”价值的关键指标，例如广告位的关键指标包括点击量、位置、大小等。最后，需求方利用这些关键指标的数据对这些事物进行排名。关于数据的需求，还有一些有趣的场景，例如，网络论坛对购买德国宝马汽车的客户的决策有多大影响力？在高排名和用户经常光顾的“男士剃须刀”网页，广告商应该购买哪个广告空间？某明星的广告推荐能达到多大的宣传作用，有多少转化成了购买力？

还有一种数据需求是收集关于某事物的事实信息，并基于事实信息建立相互联系。几十年来，这种情况在信息集成中是众所周知的，也在文本挖掘和信息检索中有着重要意义。简单地说，如果你想深入了解一件事物，那么最好的方式，就是知道它的相关数据，越多越好。不管你是用它来分析、应用，还是了解，数据都是最直接的突破口。

数据的价值需要市场里不同角色的人共同努力才能挖掘出来。了解数据市场中的需求、利益和不同受益人的需求，对数据的定价策略至关重要。图 6.2 展示了目前数据市场的基本结构。

数据市场的参与者可大体分为如下七类。

第一类是分析师。充当分析师角色的典型成员是各领域的专家，如并购专家、销售主管、产品经理、营销经理和业务分析师。这些专家最常使用的数

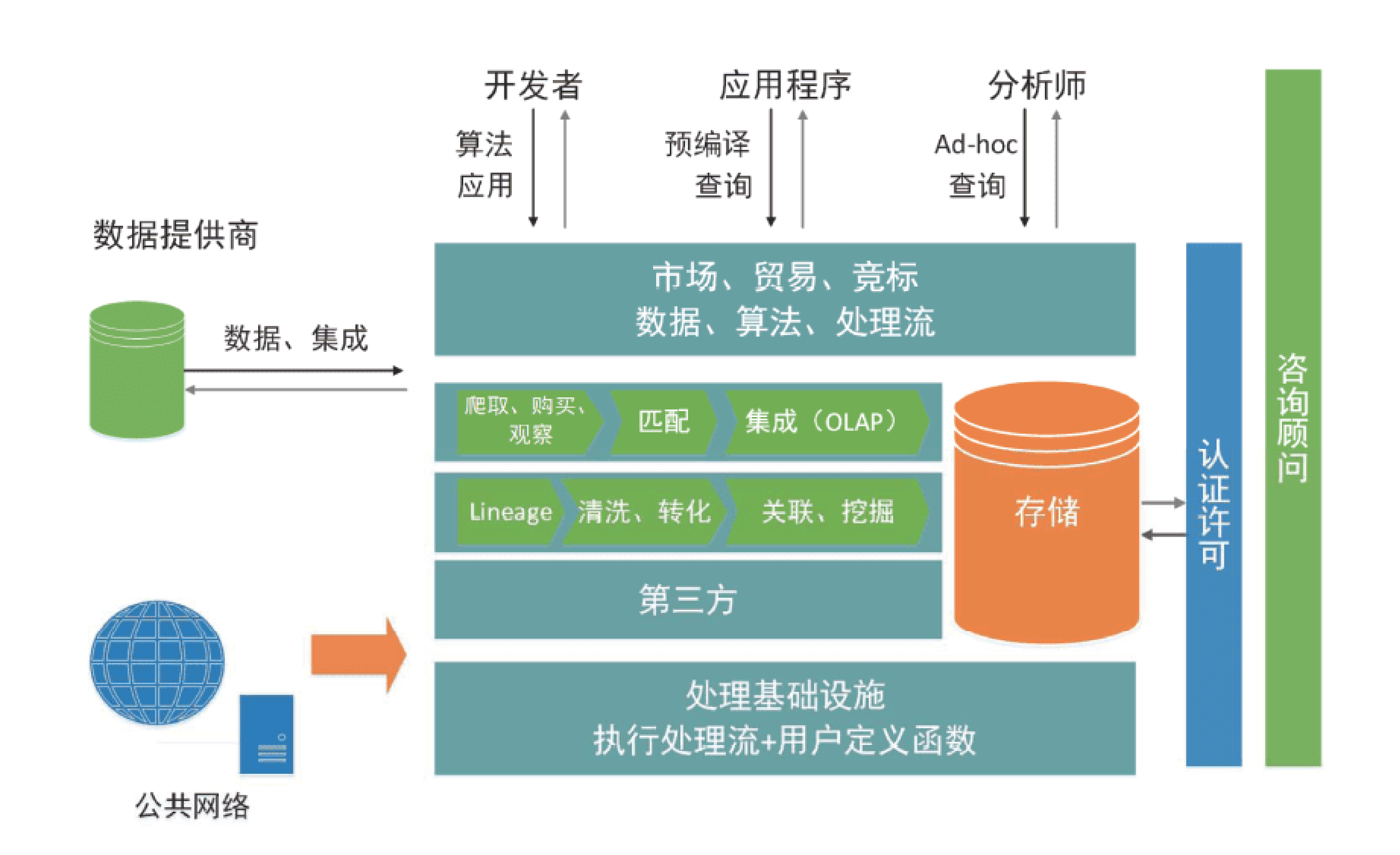


图 6.2　数据市场

据挖掘工具是 Web 搜索引擎。这些人利用办公产品（Word、Excel、PowerPoint 等）汇总数据探索的结果，形成报告，以直观、易懂的方式展示给领导或大众。从数据市场的角度来看，这个群体试图从公共可用数据（大部分来自互联网）、商业数据（商业来源）和私人数据的无限选择中受益。为此，分析师们向数据源发出即时查询，并以高度交互的方式组合数据，然后将数据源和数据集成转化成文字，变成大家都能读懂的知识。

第二类是应用程序供应商。分析师可能并不会编程，他们精湛的分析、严密的逻辑和多样的需求在计算机眼里只是毫无意义、错乱失常的电流。对数据的探索是一项包含许多繁杂而重复步骤的任务，如果让分析师来完成会极大地降低分析师的工作效率，而程序员可以将分析师的常见需求转化为应用程序。这些应用程序简化并加速了任务的进行，帮助不同领域的专家完成分析。从数据市场的定价角度来看，这种做法将会促进稳定数据和服务的产生，同时在应对连续需求上也将有所帮助，从而保障数据市场健康有序的发展。

第三类是数据关联算法的开发人员。需求方获得的数据通常来自大量的、不稳定的数据源，应用程序供应商和分析师需要集成、处理这些数据，例如，数据挖掘、匹配、清理、相关性计算和沿袭跟踪。不同的步骤会应用不同的算法。对于不同类型、不同领域的数据，这些算法的“质量”或“可加入性”是不同的，并且通常仅适合特定的领域。开发人员可以将这些算法包装成黑盒（用户定义的函数）放到数据市场，这样一来，其他参与者便可以“购买”这些算法，然后只需通过函数调用就可以实现复杂的功能了。

第四类是数据提供者。“巧妇难为无米之炊”，数据提供商的存在是数据市场得以存在的前提。这里我们将数据提供商分为商业和非商业的。商业的数据提供商既有像 Google、百度这样提供网络搜索引擎的商家，也有来自网络论坛的链接数据提供商，还有销售多年积累的金融数据和地理数据的商业数据提供商，如路透社或彭博社等。非商业的数据提供商有政府机构、联合国组织或世界银行，它们会免费提供统计数据。当然，一些数据提供商也开始提供将不同数据集成在一起的算法，兼职数据服务商。

第五类是顾问。分析师不是万能的，对于特定的专业领域，他们也需要咨询，以便于在数据源选择、集成、评估和产品开发方面做决策。这时候，顾问便可以解决特定领域的专业问题。

第六类是许可和认证实体。我们都知道买商品的时候不能买三无产品，也就是无生产日期、无质量合格证以及无生产厂家的产品。数据商品也一样。在数据市场上，我们也需要许可和认证，有人称之为“品牌标签”，用来帮助客户购买数据相关产品。

第七类是数据市场监管。如前所述，数据市场面临着技术、伦理、法律和经济方面的挑战。如何建立一个值得信赖的品牌和大型社区；如何开发用于存储、搜索、交换数据以及相关算法的公共平台，同时最小化算法执行时间；如何制定公平公正又适用于市场的法律法规，这些都是挑战。数据市场监管

的存在可以帮助参与者们从宏观的角度战胜这些挑战，使数据市场有条不紊地运转。

给数据定价是个技术活

有价值的东西只有对懂得的人才有意义。

——普劳图斯，剧作家

上面我们提到，贵阳率先成立了全国第一家大数据交易所，但是成立后发现了一个严重的问题，那就是怎么给数据定价。数据和普通的商品不太一样，它能以很低的成本复制，然后卖给很多人；并且，买家和卖家往往都不清楚它究竟会产生多少价值。中国大数据金融产业创新战略联盟常务副理事长范晓忻说："现在很多企业、政府都拥有大量数据，但这仅是'数据大'，而不是大数据。真正的大数据，是需要进行演算、评估等，而且这些数据就像未知的矿藏，存在一定风险，通过演算后可能在应用中并不存在什么价值。"这是个复杂而新颖的问题。2016 年 1 月 8 日，贵阳成立了大数据资产评估实验室，这也是全国第一家大数据资产评估实验室。

成立当天，实验室就正式发布了对人民日报媒体技术股份有限公司"传播效果评估系统"的数据资产评估成果。大数据资产评估实验室与人民日报媒体技术公司签约，将"传播效果评估系统"作为数据产品进行分析评估。评估认为，该大数据服务对于一家中等规模媒体机构每年的价值为 50 万到 80 万元，而面对全国众多机构用户，将产生亿元级的数据服务价值。

根据麦肯锡对西方产业数据的评估，大数据能使欧洲发达国家政府节省至少 2000 亿欧元的运作成本，每年给欧盟带来 400 亿欧元的经济增长；使美

国医疗保健行业降低 8%的成本;并使得大多数零售商的营业利润率提高 60%以上,帮助制造业在产品开发、组装等环节实现约 50%的成本消减。

数据为政府和产业带来了巨大的价值,那么如何对数据的价值进行衡量进而定价呢?数据是否能够按照一定的计量单位来计算每 T、每 G 数据多少钱,每一千条数据记录多少钱,每张图片、每条语音分别都多少钱?

讨论这一问题之前,我们需要清楚,这里所说的“定价”指的是确定数据的价格,而非价值。“数据如何定价”并不完全等同于“数据值多少钱”,也就是说数据的价值并不等同于它的价格。

价值也可以有“熵”有量

数据的价值是什么?是推荐一款产品,是读懂一个客户,还是计算一种人生?也许,数据价值是发现——发现你原来并不知道的问题,然后帮助你提供解决问题的方法和思路;又或许,数据的价值是自我认知。

如果这些都是数据的价值,我们又怎么去判断哪一种价值更大呢?如果数据的价值能用 1、2、3……这样的数字来表示就简单多了,这样我们就能轻松地判断孰大孰小了。

然而,真的能有这样一种方法吗?

首先,我们来考虑数据本身的价值所在。很多大数据的专家都说过:数据本身是没有价值的,真正有价值的是从数据中提取获得的信息。什么是信息呢?“张三姓张。”我们会说这句话毫无信息量。“卡文迪许姓卡。”我们会说这句话的信息是错误的。事实上,我们平常对信息的判定,是经过人脑根据信息传递目标和心理预期等因素判断的,这种判定常常是因地制宜的,人们大多考虑的是信息的意义。而从科学的角度研究信息时,这个“信息”就与我们平常所说的“信息”不一样了。

熟悉通信的人都知道克劳德·艾尔伍德·香农(Claude Elwood Shannon)。第5章介绍过香农,他是美国数学家,他在研究密码学原理时,提出了信息论。他提出:“对于信息论的研究而言,信息的‘意义’基本上无关。”也就是说,我们要剥除信息的语义内容,假装听不懂信息说的是什么,排除人心理上对信息有无意义的判定,只是“机械地”关注传来的字符、数字等。例如,“1+1=2”这条信息与“1+1=3”这条信息,它们的信息量是相同的。那么,去除掉了这些我们人类所关注的意义,信息还剩下什么了呢?

香农认为,还剩下熵。

熵又是什么?

熵,是体现混乱程度的度量,也是热力学中用来表示物质状态的参量。举例来说,看看你的衣橱,如果它很乱,你可以说你的衣橱“熵很大”;如果它很整洁,分类清楚,一尘不染,你可以说你的衣橱“熵很小”。

“熵”是从何而来呢?

“熵”这个概念并不是由香农提出来的。早在1854年,克劳修斯(T. Clausius)就提出了熵(entropie)的概念,并将其应用在热力学中。我国物理学家胡刚复于1923年首次把英文entropie译为“熵”。将“熵”字拆开来看,火表示热量,商是热能除以温度的结果。语言的魅力在于清晰且生动的表达,而翻译的魅力是在意义和风格的对等之上,更多一分奇妙的融合与创造。

熵是描述一个系统的无序程度的物理量。熵越大,无序程度越高;熵越小,无序程度越低。熵的概念提出,是为了更好地解释热力学第二定律——不可逆热力过程中熵的微增量总是大于零,又称“熵增原理”,它表明了在自然过程中,一个孤立系统的总混乱度(即“熵”)不会减小。简单地说就是,“孤立系统”只会越来越混乱,不会自己变整洁。

我们来看这张图。图6.3的左边是大自然中自然形成的沙堆,没有什么秩序,它的熵值很高;图6.3的右边是一座沙子堆砌的城堡,有规则的形状,

它的熵值很低。




图63　不同形态的沙子的熵

一切自发的不可逆过程都是从有序到无序的变化过程，向混乱度增加的方向进行，于是有了上面提到的"熵增原理"。就好比，图6.3右图中的城堡可能坍塌成一盘散沙，图6.3左图中的沙子却不会自己堆成城堡；理想气体扩散后不可能自己缩回去；在不整理的情况下，你的房间会越来越乱。要注意的是，熵增的条件是这个体系与环境没有能量交换，这个过程是不可逆过程——所以不存在"熵减"。但在生物学中，存在一个"负熵"的概念。1944年薛定谔出版了《生命是什么》一书，在该书中提出了"负熵"的概念——生物会吸收环境中的功，而减少自身的熵，因而变得有序。

人类总是善于运用思维的类比与迁移，学科的交叉借鉴形成了科学研究的良性循环。除了生物学，熵还被广泛应用于控制论、概率论、天体物理等领域。在科学研究中，熵是用来描述、表征系统不确定程度的函数；在社会科学中，熵用以借喻人类社会某些状态的程度；在传播学中，熵表示一种情境的不确定性和无组织性。后来，香农将熵的概念引入到信息论中来，用来描述信息的不确定性。

人们常说信息多、信息少，可是信息究竟有多少呢？信息熵解决了信息的量化、度量问题，我们可以用一个数字来说明这条信息有多大。它是怎么

做到的呢？

如果去掉此信息，内容的可能性越多、不确定性越大，那么此信息蕴含的信息量就越大。这里的不确定性、可能性指的就是概率。例如，在汉语中，如果紧跟在“什”后面的字是“么”，那么这其中的信息量就很小，因为“什”后面出现“么”的概率很大，不需要别人来告诉你，你也能猜到。

香农认为，信息不是创造出来的，而是选择出来的。也许你会说，这样的选择也太难了。例如，对于“算法统治世界”这样一句简单的话，每个字都是从几千、几万个字中选取出来的，这样如何计算？首先，第一个字“算”是从众多汉字中挑选出来的，但并非所有汉字接在“算”字后面的概率都是相同的，显然“法”字的概率是相对较高的，以此类推，“治”接在“统”后面，“界”接在“世”后面，概率都是比较高的，所以对于某些位置，可能出现的汉字数目是比汉字总量要少的；其次，不要忘了，在信息的世界里，计算机只认得二进制数字，所有的文字、符号、语音、图片、视频等都要转换成“0”和“1”的组合才能被计算机处理，比如字母 a 在计算机眼里是 01100001。这样每一个位置都有“0”和“1”两种可能性，计算信息的不确定性便有迹可循了。

每个符号传递的信息量与可能符号的出现概率有关。香农用下面这个式子告诉我们怎么算信息量的大小。依据式子计算得到的信息量的单位是比特，因为式子中的 2 代表的是二进制。

$$H = -\sum p_i \log_2 p_i$$

初学“熵”时，我们会用某种具象化的东西来帮助自己理解“熵”；等我们真正弄懂了“熵”，我们又不得不把这些乱七八糟的东西从脑子里清除出去。

这就是熵。

回归到数据的价值上来。如果说数据的价值是信息，那么这里的信息又是指什么呢？在数据交易的场景下，我们考虑的是对交易产生价值的信息。

在此，我们提出两种对价值熵的定义。

一种是根据数据的不确定性来定义。对于数据的各个字段，如性别、年龄和身高，如果每个字段的取值可能性多，则其熵大，对各个字段的熵进行加权，最终获得的总熵便是这份数据的价值熵了。价值熵越大，则数据的价值越大。

另一种定义是利用数据各个字段之间的相关性。数据各个字段之间的相关性（包括正相关和负相关）强，且相关关系数目多，数据的价值熵小，其价值更高。

这里仅仅对价值熵进行了简要的介绍，价值熵的定义和应用仍然有待进一步的深入研究。

价格多少谁说了算

什么是数据？

123，这是数；一场电影，两段对话，三种心情，这也是数据——以数为依据，描述一个实体，或者一个事件。

什么是交易？

李逵会砍柴，张顺能打鱼。李逵用自己砍的柴换张顺打的鱼。这是交易——后来这样做的人多了，柴太重，鱼不好拎，于是用一样东西代替柴和鱼去交换，那就是钱。

对大部分人来说，一条鱼能顶半天的饱，相当于 1 斤柴的价格，值 10 块钱；然而，一份数据等价于多少斤柴，值多少钱，答案对每个人来说都是大相径庭的。

在第 4 章中，已经介绍了许多传统的定价方法，但由于数据商品的特殊性，传统的定价方法并不适用于数据定价。我们需要提出新的数据定价算

法。我们将把数据定价分为两个过程：数据价值评估和定价策略。前者决定了某份数据最多能卖多少钱，最少可以卖多少钱；后者决定了用什么样的策略来决定当前它具体该卖多少钱。听上去有点绕，其实就是商品的成本和价格的关系。

卖方确认数据的价格区间

首先，交易者要确定价格区间，也就是某个数据商品的最低价格和最高价格。我们可以用成本法来估计数据商品的最低价格，例如，根据数据提供商付出了多少人力物力，以及他们本身对数据的基本估值，可以推算出数据的最低价格。也许你会问，确定最低商品价格就好了，为什么还要给它一个上限呢？难道有钱不赚？

事实上，当商品的成交价超出了估计的最高价格，很可能影响数据交易市场的稳定性，也不排除一些非常、甚至非法的市场运作。为了数据市场能够健康的发展，最高价格的估计是有必要的。

卖方出价

确定好价格区间后，就要进行精确定价。正如莫里克所说，事物只有当人们认为它们有价值时，才有价值。对于不同的需求方来说，同一份数据价值很可能相距甚远，所以如果不考虑需求方，直接制定数据的价格是没有意义的。在利用基本的价值定义确定数据的大概价格区间后，最重要的还是利用定价策略来对每一份数据进行精准地定价。

卖方并不知道每个需求方愿意出多少钱购买他/她要出售的数据商品，最好的方法就是让他们自己说出来。

这就像拍卖一样，也是我们认为目前最适合作为数据定价策略的方式：让需求方自己出价。

说起拍卖，大家一定会想到影视剧中的商战现场。拍卖者拿出一件物品，多位衣冠楚楚的竞拍者坐在下面，有的势在必得，有的神情紧张，有的露出暗黑式的笑容，有的用眼神与对方先来了个刀光剑影。只待拍卖者先说出起价，随后一场暗潮汹涌的竞拍便拉开了大幕。

其实，拍卖远比电视剧中演得复杂。拍卖是经济学研究的一个重要问题。经济学界普遍认为：拍卖是一个集体（拍卖群体）决定价格及其分配的过程。

拍卖有三个基本条件：

首先，拍卖必须有两个以上的买主（多个买主才能产生竞争）；

其次，拍卖必须有不断变动的价格：拍卖并不是买卖双方简单地讨价还价，而是由买主以卖主当场公布的起始价为基准另行报价，直到没有人再加价为止；

最后，拍卖必须有公开竞争的行为：拍卖是不同的买主在公开场合竞相出价，竞争同一物品。如果所有买主都无心竞价，没有了竞争，拍卖就失去了意义。

拍卖的方式也有很多种。大家最为熟悉的便是上述影视剧里最经典的拍卖，它的学名叫做英格兰式拍卖，也叫增价拍卖。顾名思义，就是竞拍者根据拍卖者给出的起价开始向上叫价，直至拍卖者愿意出的最高价出现，拍卖最终成交。

有增价拍卖，就有降价拍卖。降价拍卖又叫荷兰式拍卖。降价式拍卖通常从非常高的价格开始，价格太高时没有人竞价，这时，价格就以事先确定的数量下降，直到有竞买人愿意接受为止。降价式拍卖的第一个实际的竞价常常是最后的竞价。一旦有人竞价，就成交。虽然只有一个竞价，却也是有竞争存在，能反映出竞拍者的预期——若此时不出价，物品就会被别人拍走。当然，降价拍卖更多地应用在拍卖物品的品质参差不齐的情况下。第一个出

价最高的竞买人可以买走全部物品，但往往只买走这些物品中最好的，然后拍卖继续，价格下降，当另有竞买人愿意接受竞价，他也有同样的选择，也是买走余下中最好的，然后拍卖又继续。在这种情况下，虽然竞买人大部分时间都沉默不语，但是在竞买者之间确实存在持续的竞争。

还有一种影视剧中的常常出现的拍卖方式，叫密封递价式拍卖，又称招标式拍卖。这种拍卖经常出现在土地使用权、较大的库存物资的竞争中。由竞拍者在规定的时间内将密封的报价单（也称标书）递交拍卖人，然后拍卖人选择竞拍者。这种拍卖方式与上述两种方式相比较，有两个特点：一是除价格条件外，还可能有其他交易条件需要考虑；二是可以采取公开开标方式，也可以采取不公开开标方式。

但是，普通的拍卖并不能够满足数据交易的特性。就如之前提到的，数据作为商品，是可以复制的，也就是说，这场拍卖的得标者可以不止一个人，而对于需求方来说，获得这份数据的人多了，也有可能对自己产生影响；同时，数据又是可以重复拍卖的、有时效性的，随着时间的流逝，数据的价值是会发生变化的。所以参与竞争的需求方需要向交易平台提供的信息除了出价以外，还应该提供自己能够接受最多与几位竞标者共同得标，以及距离下次拍卖同类数据至少间隔多长时间。对于数据平台来说，如果多人得标，如何对每一位得标者收取费用也是个棘手的问题。

数据定价是一个新问题，看上去我们之前没有遇到过，但这并不代表我们对其束手无策。下面我们借助几块“他山之石”，希望能找到解决数据定价问题的新思路。

从没有拍卖锤的拍卖到数据的拍卖

我们印象中的传统拍卖场总是惊心动魄，暗潮汹涌。1766 年，一位来自

澳大利亚帕斯的苏格兰人在伦敦开设了世界上第一家艺术品拍卖行，他的名字叫詹姆士·佳士得(James Christie)，于是拍卖行的名称就叫“佳士得拍卖行”(最早也翻译成克里斯蒂拍卖行)。图 6.4 是佳士得拍卖行的拍卖场景。



图 64　传统拍卖会

实际上，有许多拍卖并没有拍卖锤，也没有拍卖师，没有举牌，甚至没有拍卖场，有的只是一台计算机和一段程序，整个拍卖周期不过几秒钟。在赛博经济世界中，拍卖是一种算法，是由计算机执行的、一种程序化的交易方式。这种算法的正确性体现在它是否符合拍卖规则，它的效率由拍卖完成时间决定，而它的性能则由拍卖者的收益决定，图 6.5 给出了拍卖与算法的映射关系。

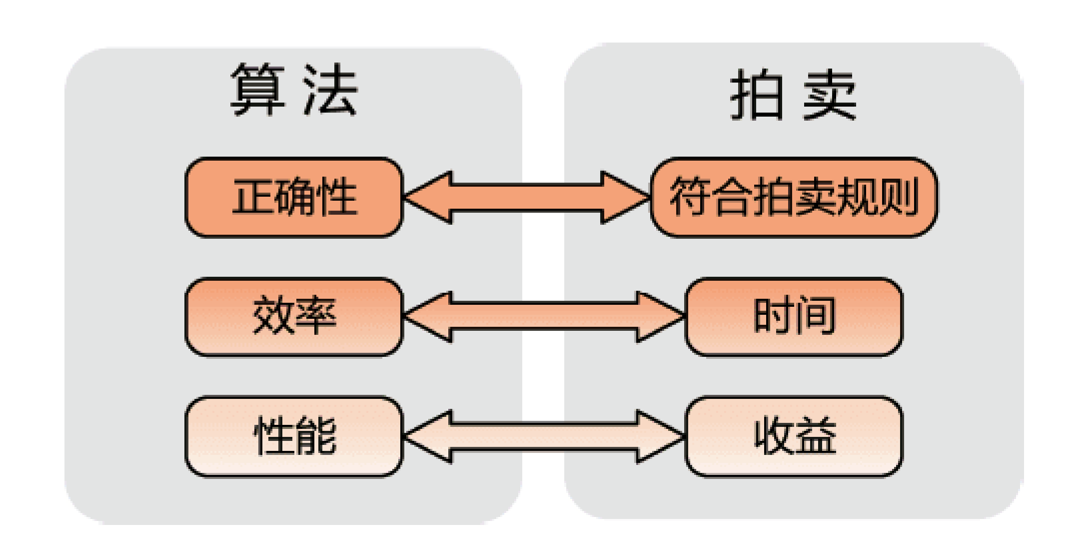


图 65　拍卖与算法的映射

我们用两种典型的没有拍卖锤的拍卖来看看数据交易价格是如何确定的。

广告拍卖的做法

通过第5章的介绍，读者已经对搜索算法有所了解。但当我们真正用搜索引擎进行搜索时，看到的第一条搜索结果往往并不是根据前面介绍的搜索算法算出来的，而是“广告”。并且，细心的朋友会发现这样一个现象，即使使用同一个搜索引擎，搜索同样的内容，出现的广告也可能不同。图6.6是笔者10分钟内两次输入“大数据”，某搜索引擎显示的结果：在第一次搜索中，某在职大数据机构成功拔得头筹，腾讯大数据分析平台斩获亚军；在第二次搜索中，阿里云折桂，而腾讯以稳定的发挥保住了第二。由此可见，这场交易并不像传统广告那样，广告主给了地铁广告费，他们的产品就会按照规定的时间出现在地铁上。

这是一场实时交易，这是一场持久战，这是一场博弈。

假设某网站有一个广告位要出售，而有10个广告主看上了它。怎么办？卖给甲？乙给的钱多；卖给乙？丙的点击率高；卖给丙？丁和丙难分伯仲……如果能雨露均沾就好了。于是，一个雨露均沾的办法——实时竞价出现了。

广告主可以像优化自己的推荐系统一样优化广告。简单地说，假如你最近在淘宝中搜索过连衣裙，当你再次打开淘宝的时候，它会向你推荐各种各样的连衣裙，这是淘宝内部的推荐，这里用的算法是推荐算法；然而这时，如果你打开百度浏览新闻就会发现，淘宝的连衣裙店铺们出现在了搜索结果的第一位上，那么这一次就是广告了。若此时，你通过百度搜索的结果点进了淘宝网，那么淘宝就会为你的这次点击给百度付费。淘宝的这次出现，背后是一个广告系统的运作，而决定它冲破囚栏出现在你面前的正是竞价广告计价算法。

其实，对这次展示虎视眈眈的商家可不只淘宝，还有唯品会、京东、当当等电商平台。这些商家都在广告平台上注册并填写了如图6.7这样的表格，



图 66　搜索广告示例

说明自己的目标用户和出价。

每隔一段时间，广告商家会进行一次投标竞价。当然，这并不是广告主派了几个员工去点鼠标、敲键盘竞价，而是已经得知了广告主需求的广告平台自动运行程序，对每个空闲的广告位进行拍卖。

没错，互联网广告的竞价方式就是拍卖。



图 6.7　互联网广告投放的定向条件

那么，在计算广告系统中，拍卖方式又是怎样的，媒体网站该向赢得广告位的广告主们收多少费用呢？最简单的方法就是按照它之前的出价进行收费。开始的时候，谷歌和雅虎都是这么做的。我们称这种拍卖方式为 GFP（Generalized First-Price，广义一阶价格拍卖），如图 6.8 所示，在该图中，广告主 A 的出价最高，因此获得广告位 1 的使用权，广告主 B 的出价第二高，因此获得广告位 2 的使用权，广告主 C 出价最低，所以拿不到广告位，广告位 1 每小时发生 100 次点击，广告主 A 会为此支付 100 元，同样，广告主 B 会为广告位 2 支付 25 元。

但是，他们很快发现了问题——这种做法很可能导致系统不稳定。具体地说，就是提供广告位的网站收益不稳定，竞价过程效率低。例如，拍卖开始时，广告主 A 出价 0.1 元，见状，广告主 B 会出价 0.2 元，广告主 C 退出竞争。接着，广告主 A 出价 0.3 元，广告主 B 出价 0.4 元……我们看到，价格会

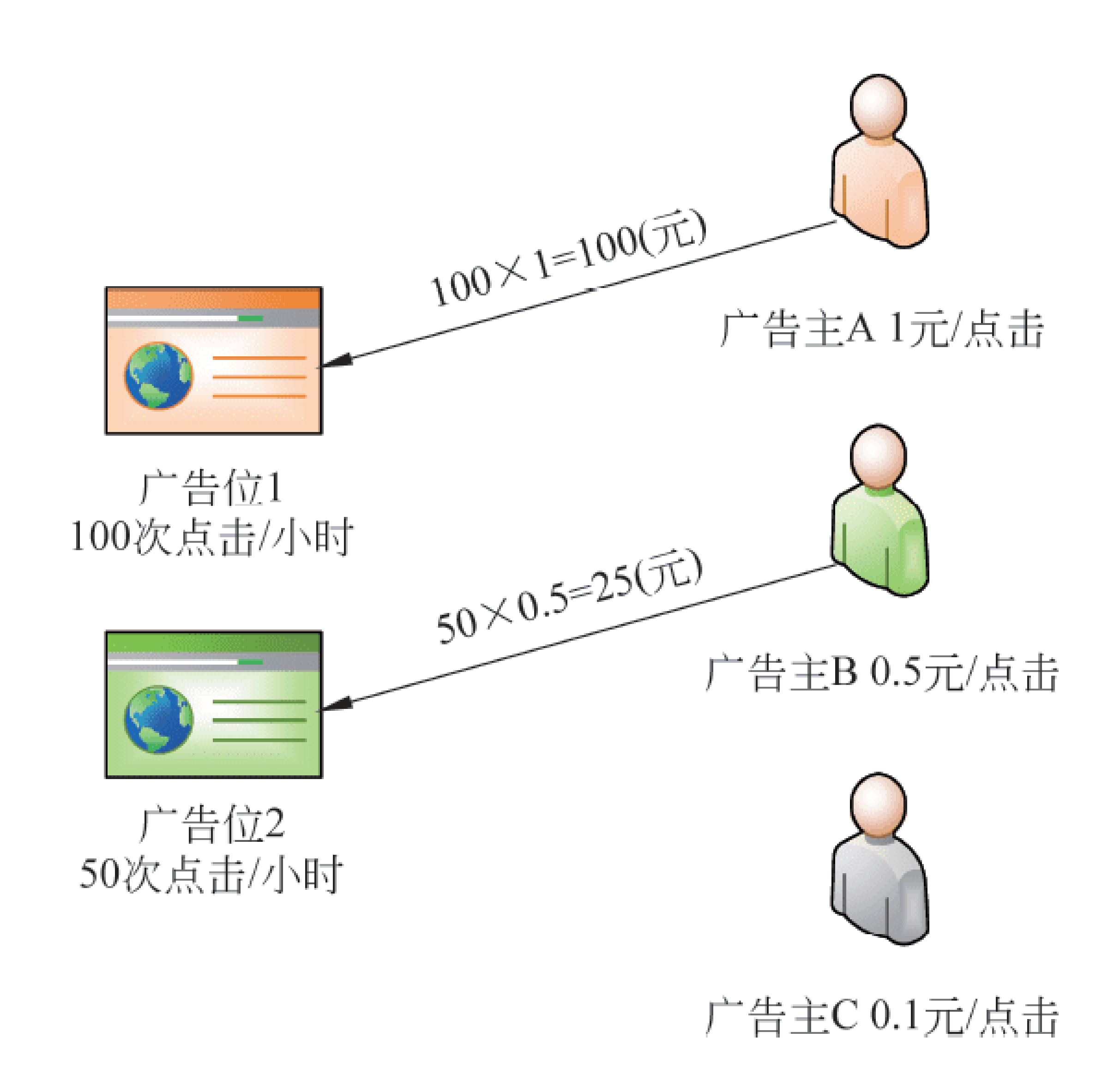


图 68　广义一阶价格拍卖示例

以一个非常小的差距“稳步上升”，直到广告主 A 出价 0.5 元的那一刻，广告主 B 决定离开。因为广告主 B 认为这个广告位的每次点击只值 0.5 元，也就是只能给他带来至多 0.5 元的收益，所以当价格超过 0.5 元时他退出了竞争。而在下一次竞拍这一广告位的时候，A 想，我都没有竞争者了，我还出那么高价格干吗？于是 A 降低了出价。而 B 得知这个消息后，又回来加入到竞拍中，于是微小差值的螺旋竞争再次上演。然而，这种价格的反复波动显然不是市场想见到的。大家都希望有一个机制能够使市场稳定下来。所谓稳定，是指整个系统处于均衡状态，简单地说，就是对于最终赢得这若干个广告位的每一条广告，它的收益都比排在其他位置上高。每个广告主都达到了自己希望的收益最大状态，系统自然就稳定了。

2002 年，Google 提出了 GSP（Generalized second-price，广义第二高价）竞价算法。2007 年，哈佛大学的 Benjamin Edelman 等在 *The American Economic Review* 发表文章详细分析了 GSP 算法。

直到今天，在真正的广告竞价系统中，大多数提供广告服务的媒体网站

依然在使用 GSP 来为广告位定价。下面我们就来讲讲这个主流竞价算法——GSP。

在广告位拍卖之前，媒体网站先要知道广告在广告位上的点击率。点击率表示的是：用户到达网站，看到了广告，有多大概率会点击这条广告。从宏观的角度，点击率又可以等价于在所有进入网站，看到这条广告的人中，有多少人点击了广告。显然，在广告拍卖的整个过程中，点击率都是未知的。所以，我们需要预测点击率。如何预测呢？当然不是掐指一算。那么是按照第 5 章中搜索里的排序算法对各个广告位进行排序吗？也不行。搜索排序的最终目的是使搜索结果的排列顺序正确，而广告系统需要的是对点击率本身的预测正确。为什么广告系统需要知道点击率的具体数值呢？这就要说到拍卖过程中一个非常重要的环节——对候选者进行排序。广告系统对各位候选广告主的排序，不单单是根据广告主的出价，而是根据 eCPM (effective cost per mille)进行排序。所谓 eCPM，是指千次展示收益的估值。按照用户点击广告和转化成购买行为这两个发生在不同阶段的行为，eCPM 可以分解成点击率和点击价值（单次点击为广告产品带来的收益）的乘积。对于媒体网站来说，它可以通过计算 eCPM 来预估广告位的效果如何。所以在这里，点击率预测不是一个排序问题，而被看成一个回归问题。

建立点击率预测模型，就是找出点击率与广告、用户与环境之间的关系。虽然只用三个词“广告”“用户”和“环境”就概括了我们需要的特征。但实际上它们对应的具体特征数量巨大，例如，广告可以由广告主、广告计划、广告组、广告创意等多层次的标签组成，用户和环境又有其各自的标签，而它们的组合数量可想而知。系统对表现这些特征的海量数据进行在线的机器学习并非易事。于是，聪明的工程师想出了一些方法来降低组合数量，简化运算。在建立模型之后，会用迭代的方法进行优化。同样，迭代次数也成为计算机的眼中钉，因为实在太多了。于是，工程师又想出了一些方法来减少迭代次

数。当然，这些方法都是科学合理的，为的是提高计算机解决问题的效率。

得到了点击率的预测值后，拍卖流程正式启动。为了便于解释 GSP 这个概念，在这里，我们假设点击率只与广告位有关。在下一节中，我们会讲到在真实情况中由谁来估计点击率，根据什么估计。

言归正传，GSP 到底是什么意思呢？它的中心思想就是，对于赢得每个位置的广告主，都按照排在他下一位的广告出价来收费。

我们看图 6.9，假设已经预测出广告位 1 每小时有 100 次的点击，广告位 2 每小时有 50 次点击。广告主 A 每次点击出价 1 元，广告主 B 每次点击出价 0.5 元，广告主 C 出价 0.1 元。三人开始竞拍，广告主 A 得到广告位 1，但他只需要付 100×0.5＝50(元)，其中 0.5 是广告主 B 的出价。同样地，广告主 B 得到广告位 2，他最终付的钱为 50×0.1＝5(元)，其中 0.1 是广告主 C 的出价。在这样的策略中，对于每个广告位，出价最高的广告主，只需要付第二名的出价。我们可以这样理解 GSP，A 的出价最高，说明他最看重这个广告位，所以我们把广告位 1 给 A，同时我们非常仁义，只需要 A 出 B 的报价即可。这样 A 就以比自己预算低的价格拿到了广告位 1。

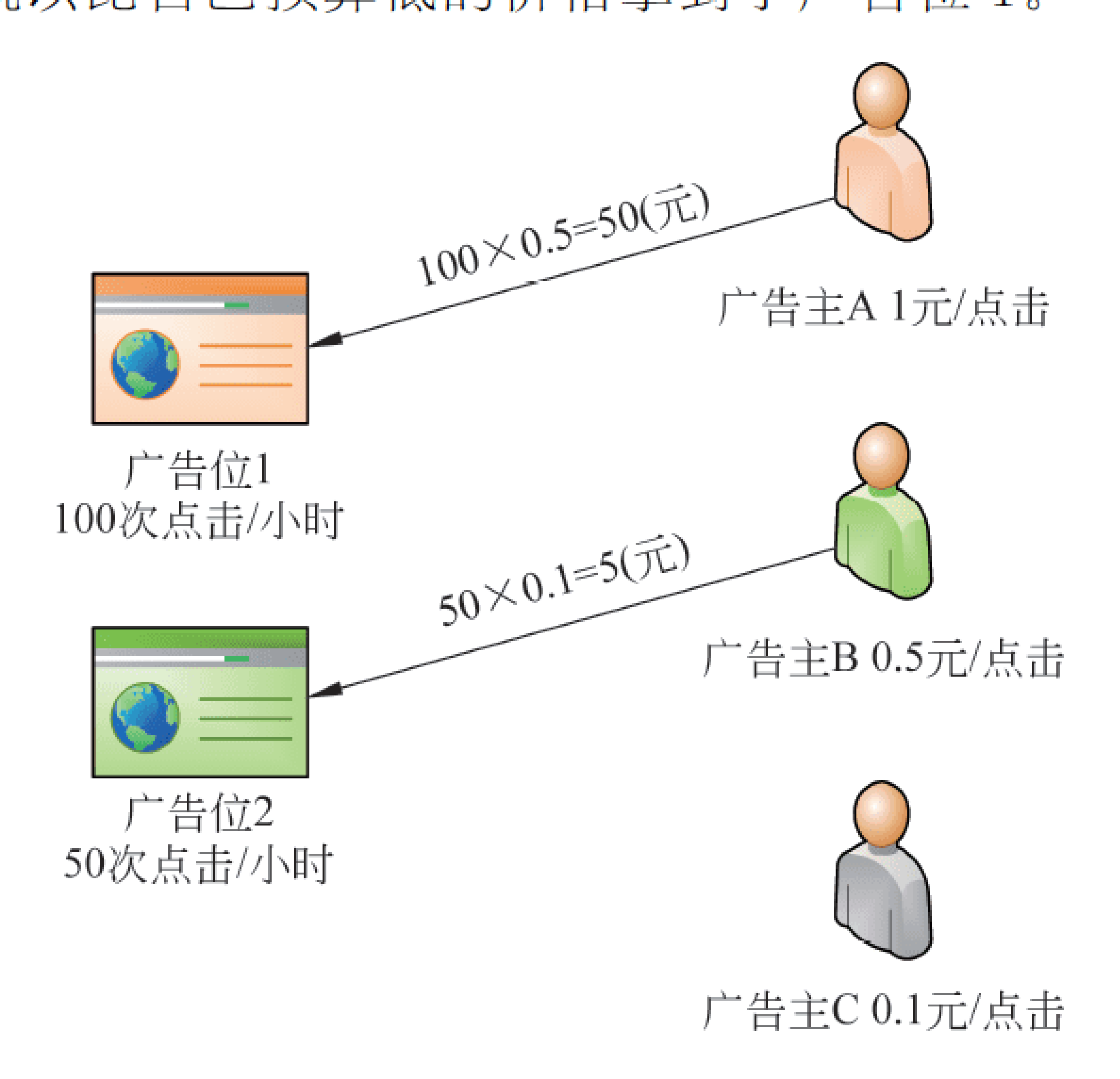


图 69　广义第二高价拍卖示例

虽然 GSP 是最为主流的定价策略，但它并不是广告实时竞价的最佳策略。

GSP 存在的一个主要问题是，广告主按照自己对广告位的真实估值报价并不是最优策略。大家可以设想一下，如果广告主的最优策略是按照自己的真实估值报价，事情会简单很多，广告主不用处心积虑地考虑如何报价来获得最大收益了，直接报自己的真实估值就行了。那么是否存在可以鼓励广告主按照真实估值报价的机制呢？

好消息是这样的理论最佳策略确实存在，VCG（Vickrey-Clark-Groves）机制就是典型的代表。

VCG 的核心思想是：对于赢得广告位的广告主，其所付出的成本应该等于他占据这个位置给其他市场参与者带来的价值损害。简单地说，他需要付的费用，就是其他人的损失。那么如何计算其他人的损失呢？很简单，把他去掉，然后看其他人的收益增加了多少。我们来看图 6.10 所示的示例。

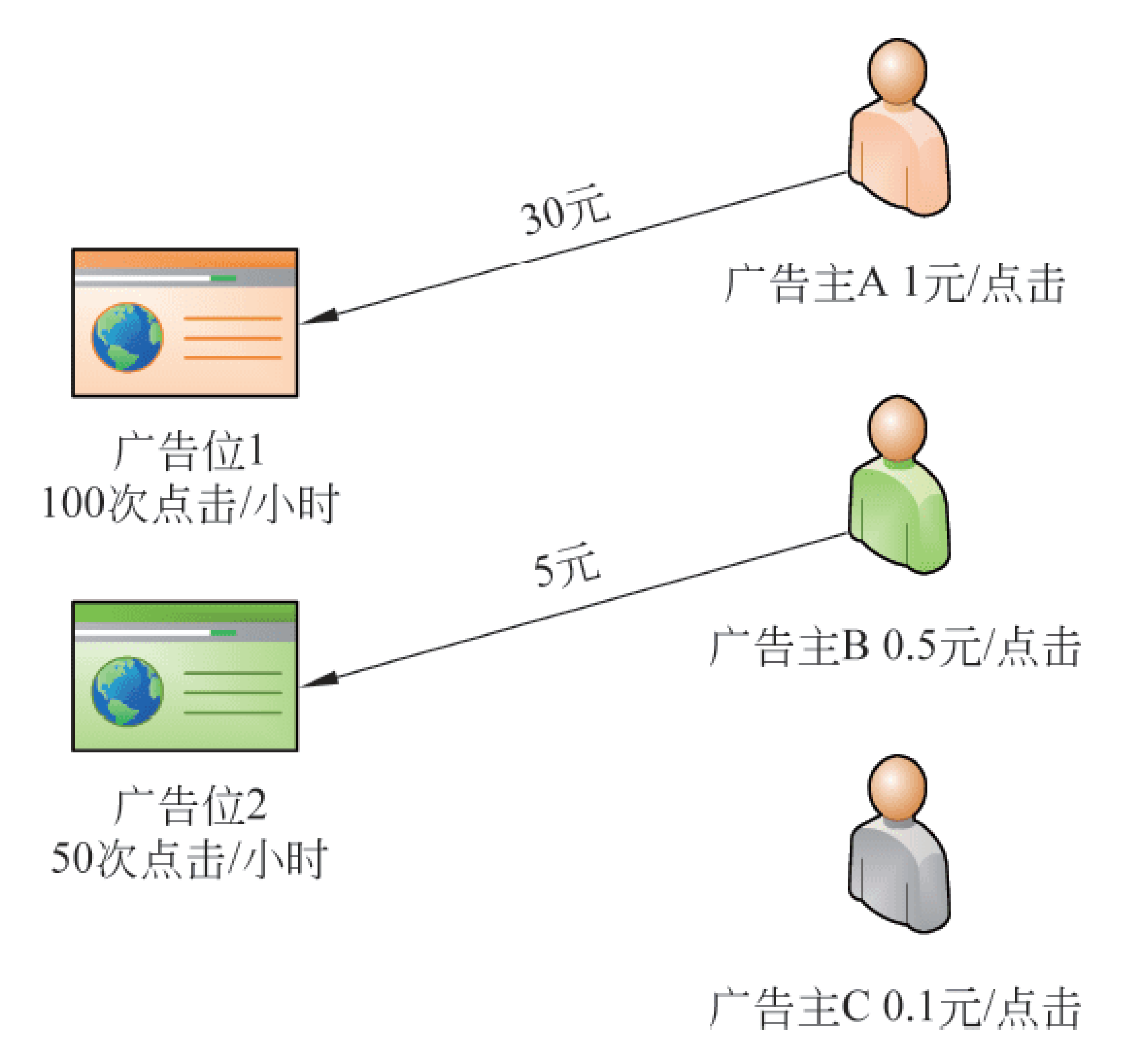


图 6.10　VCG 拍卖示例

广告主 A 每次点击出价 1 元，广告主 B 每次点击出价 0.5 元，广告主 C 出价 0.1 元。现在只有两个广告位，第一个广告位平均每小时有 100 次点击，第二个有 50 次点击。A 得到广告位 1，B 得到广告位 2。在广义二阶价格拍卖中，A 要付费 50 元，B 要付费 5 元。如果使用 VCG 方式来定价，必须考虑如果 A 没有参与拍卖会发生什么。如果 A 没参与，那么 B 会得到广告位 1，会再多得到 50 次点击，并且 B 认为每次点击值 0.5 元，所以 A 的出现意味着 B 损失了 25 元，此时，C 会得到 50 次点击，C 认为每次点击值 0.1 元，所以 A 出现意味着 C 损失了 5 元。根据 VCG，A 的付费应该是给 B 和 C 带来的损失之和，即 25＋5＝30 元。然后，考虑如果 B 没有出现的情况，此时，A 的位置不会受到影响，C 则会因此每小时失去了 50 次点击，对 C 来说损失是 5 元，所以 B 的 VCG 付费应该是 5 元。

通俗地说，鼓励广告主按照真实估值报价就是鼓励广告主讲真话，为什么 VCG 机制有这样的效果呢？详细的理论证明我们不过多介绍了，大家只需要记住一点，这里的关键是一个广告主该付多少钱是由其他广告主的出价决定的，在这种情况下，广告主最佳策略就是讲真话，因为说假话也不会带来什么好处。

VCG（拍卖机制）的全称是 Vickrey、Clarke、Groves 三个人名字的缩写。在之前传统拍卖的介绍中，我们已经了解过 Vickrey 拍卖。不难猜出，Clarke 和 Groves 两个人在 Vickrey 的基础上研究出了 VCG 拍卖机制。然而，在闻名的 VCG 中，只有 Vickrey 获得了诺贝尔经济学奖。这是为什么呢？

回忆前面介绍的 Shapley，我想聪明的你已经有了答案。没错，历史总是这么惊人的相似。Vickrey 确实是 VCG 之一，获诺贝尔奖的也是他。但使他获得诺贝尔奖的，却不是 VCG 这一成果，而是他在信息经济学、激励理论、博弈论等方面做出的重大贡献。

1935 年，Vickrey 获得了耶鲁大学理学学士学位。1937 年获哥伦比亚大

学硕士学位。1945 年起，Vickrey 开始留校任职。1947 年又获哥伦比亚大学哲学博士学位。任何一门学科学到极致，便是哲学，可见 Vickrey 的学术造诣是非常之高的。而后，他开始在经济学上有所见长，这一发展起来，便一步一步走向了诺贝尔奖。这期间，他又在 1979 年获得了芝加哥大学的人文学博士学位。

看过他的教育经历，大家可能会猜想：理学出身、一直在学校里没有见过外面世界的 Vickrey 应该是个理论经济学家吧。然而，事实总是超乎想象的。Vickrey 是一位典型的应用经济理论家，他的研究过程你可能不懂，但他研究的东西都是大众关心的、接地气的东西。

Vickrey 的第一个最富于创造性的成果，是他对税制结构方面的研究。Vickrey 想要设计一种最优的税收体系，尽最大可能平衡公平与效率。他认为如果只考虑公平而不同时考虑调动积极性，收税人员就会从富人那里收取税金，将其中一部分再分配给穷人，一直到将富人的税收提高到他们认为公平的水平，最后使人们的税后收入大体相等。显然，这种“劫富济贫”的方法不仅不能激励人们发挥自己潜在的劳动生产能力，反而会促使人们隐瞒他们的实际能力。因为具有生产力高的工人能够挣得更多的收入，但他们努力获得的更高收入会被更高的税率所征收。所以，在这样一个纯粹是追求公平的税收体系下，最具有生产效率的工人将不会十分卖力地工作。于是，Vickrey 提出了 21 条改革美国所得税体系的建议。他发明了“累积平均制”“遗产权继承税制”，Vickrey 还对消费税、公司税、政府债券的税收减免、土地价值税等方面有许多研究。

20 世纪 60 年代，Vickrey 开始对拍卖等具体的市场机制进行研究。投标或喊价有着悠久的历史。传统观点认为，如果交易者双方所掌握的信息是不对称的，那么市场上产生的均衡结果将是一种无效率的状态。Vickrey 却证明并非必定如此。拍卖市场是否有效率，取决于拍卖规则是否能够保证没

有人可以通过损害集体利益去实现自己利益的最大化，能不能有效地诱导参与者主动说出他们真正愿意支付的价格。首先，Vickrey对已有的标准的拍卖进行了分类，形成了一套完整的拍卖与投标理论。然后，他开始了更深入的研究。Vickrey首次研究了密封投标拍卖问题，并且对市场激励机制与信息之间的关系进行了开拓性的探索。他强调市场规则的制定必然要受到激励一致性的约束，其中市场激励是从交易者的私人信息和交易者参加或不参加交易的选择自由中诱导出来的。

Vickrey对于投标的研究，其重要性不只局限于投标方面。因为投标方法解决的是，如何在信息不完整或其分配不对称的情况下，最有效地配置资源的问题，这开创了信息经济学研究的先河。Vickrey对投标与喊价的研究，带动了许多相关的研究，诸如保险市场、信用市场、工资结构等问题。

金子总会发光，为人们做了这么多实事的Vickrey，终于获得了属于他的荣誉。只是这个荣誉来的时间，有点晚。

1996年10月8日，瑞典皇家科学院决定把该年度的诺贝尔经济学奖授予英国剑桥大学的Mirrlees与美国哥伦比亚大学的Vickrey。

在得奖三天之后，还未来得及参加颁奖典礼，Vickrey去世了。有人说，诺贝尔的光环为其人生画上了一个完美的句号。

数字商品是怎么卖的

电子书、飞机上的电影、手机上的App等商品有一个共同的名字，叫做数字商品。数字商品和数据商品有着许多相似的地方。它们都没有实体，都可以接近零的边际成本进行复制，还都可以快速“运输”到顾客手中。数字商品的定价对数据定价是有启发的。在电商网站上，我们能看到许多数字商品的买卖，它们大多是明码标价。

可是，虽然数字商品可以几乎不需要成本地复制无限份，但制定一个最

优的价格是个很大的难题。商家算不准有多少人想买，也很难估计他们愿意付多少钱。而且有时候，商家并不想无期限、无限制地供应一本电子书或飞机上播放的电影。

如今，数字商品的购买、支付、物流都可以在线上“瞬间”完成，那么如果定价也可以在线上、互动式地、“瞬间”完成，岂不省去了“砍价”“犹豫”“价比三家”的过程，大大提高了买卖的效率？

线上拍卖刚好可以解决这个问题，但数字商品的线上拍卖和广告位的线上拍卖有所不同。在广告位的线上拍卖中，一个广告位只能投放一则广告，有一个中标者；在数字商品的线上拍卖中，一份商品的中标者不止一个，成交价也不止一个。那么问题来了：在一次拍卖中，同一种商品卖出多少份最合适？对每位中标者收多少钱才能获得最大利润？怎样才能引导竞标者说出自己愿意支付的最高价格呢？

要寻求以上问题的最佳答案，拍卖算法的设计尤为重要。每位竞标者都有一个对商品的心理估值，研究拍卖的人会将拍卖者利益最大化作为目标，用能让他们从竞标者兜里掏出最多钱的方法来计算拍卖结果。算法不怕麻烦，不怕计算量大，因为计算机程序可以解决这些问题；但是，如果算法太过烦琐，不好向竞标者解释，万一解释不清，导致买家不敢轻易下手，就得不偿失了。所以拍卖算法的设计要加入商业性的因素，而不是简单的最优化问题。

不同的拍卖算法有不同的性质，下面我们对比一下确定拍卖算法和随机拍卖算法。确定拍卖算法，就是在计算有多少人中标以及对应成交价的时候，每个步骤所用的方法都不含随机性；而在随机拍卖算法中，计算拍卖结果的步骤有随机的过程。设想这样一个场景，对于投入拍卖系统的每一个标，我们用其他的标计算出一个参考价格，如果这个标超过了参考价格，那么它中标，否则竞标失败。在计算参考价格的过程中，我们可以列出一个公式，用

除这个标以外的系统内所有其他的标来计算，比如取其他所有标的平均值，这就是确定拍卖。如果我们先在其余所有标中随机选出一半的标，再用这一半的标的平均值作为参考价格，这就是随机拍卖。

也许读者会问，为什么在计算参考价格的时候要除去这个标本身呢？答案是，为了让竞标者说真话。理论推导证明，标独立的拍卖是可以让竞标者“讲真话”的。标独立的意思是，竞标者投递的标只决定自己是否中标，不决定其中标后的成交价。也就是说，你中标了以后，交的钱数是由其他竞标者的标计算出来的，并且它不会超过你投递的标。

因为拍卖者的目的不同，所面对的买家也不同，所以要因地制宜地设计拍卖算法，才能达到目标。

数据该怎么拍卖

从荷兰拍卖、英式拍卖，到广告位拍卖、数字商品拍卖，都是竞标者向拍卖者报出自己的出价。拍卖者先给所有出价排个序，然后根据事先设定好的拍卖规则，选择一个或者多个出价高的人中标，最后宣布结果，将物品交给中标的人。也就是说，在传统拍卖中，出价高的一定会中标。

然而，出了高价的中标者真的愿意与别人共享一份数据吗？在他们的心里，也会有很多顾虑与犹豫——我出多少钱才能中标？我出 10 块，会与几个人共享，我出 5 块，又会与几个人共享呢？对于企业和组织机构来说，数据的共享情况很可能直接影响数据为它们带来的收益。

例如，有一份视频用户的数据，如果独家卖给爱奇艺，那么爱奇艺会获得 1000 万用户；如果独家卖给腾讯视频，则腾讯视频将获得 1000 万用户；如果同时卖给爱奇艺和腾讯视频，则两家会分获 500 万用户；以此类推，如果同时卖给 4 家视频网站，则每家获得 250 万用户。如果进行数据拍卖，爱奇艺最

多能接受与一家视频网站共同获得视频数据，腾讯视频最多能接受与两家共享。

假设有 10 家企业共同竞标一份数据，根据出价和利润计算，拍卖者让爱奇艺、腾讯和优酷土豆三家共同中标了，此时爱奇艺陷入了进退两难的境地。如果它退出，就会影响拍卖中的所有参与者，扰乱秩序，而如果爱奇艺勉为其难地接受了三家共享，那么它的收益将大打折扣。在下一次拍卖中，爱奇艺很可能说出一个较之前低很多的价格，因为它不知道它将与多少人共享这份数据。

如何能周到地考虑各方利益，使拍卖结果出现后，各个中标者不会有“退标”的冲动呢？

针对这个问题，我们提出一种新的拍卖算法，让竞标者说出愿意支付的最高价格的同时，也说出最多愿意与几人共享一份数据。如图 6.11 所示，甲、乙、丙、丁四人投标，如果选择出价最高的人中标，那么甲当之无愧，并且按照甲的要求，只能他自己中标，此时拍卖者可以获得 15 元收益；按照四位竞标者对中标人数的要求，我们发现可以选两个人共同中标，那么有乙丙、乙丁或丙丁三种可能，显然，在这三种组合中，能让拍卖者获利最大的是乙丙组

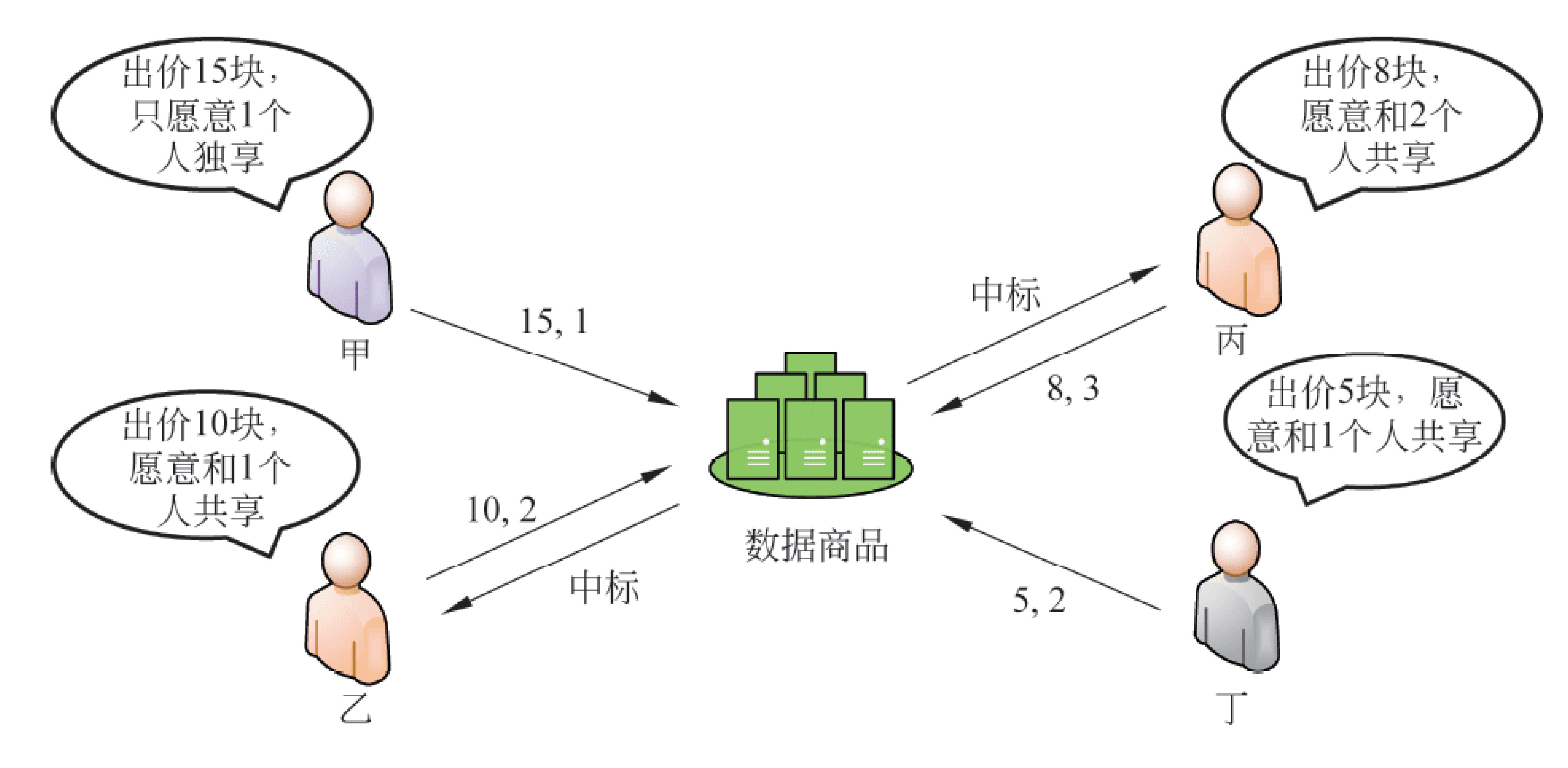


图 6.11　分布式分配拍卖

合，此时如果对两人都收取 8 块钱，那么拍卖者可以获得 16 元；而出价最低的竞标者丁无论在任何情况下都没办法中标。这时候，拍卖者很容易计算出哪一种中标结果能够获得最大利润。

在这个新的拍卖里，竞标者需要“喊出”两个数字——出价和最多能接受与多少人一起分享数据。这样一来，出价最高的人未必能中标，但中了标的人都能够接受与其他中标者共享这份数据。竞标者对拍卖结果有所约束。拍卖者在做中标人数的决策时，也可以少一些不确定的猜测，真正做到最大化自己的收益就可以了。

当然，我们可以看到，这种拍卖模式并不适合在会场里以举牌叫价这种方式进行，虽然每个竞标者只是多“喊”了一个数字，但带来的运算却不止一步。在下一节我们会给出一种在线的数据交易场景，让拍卖以算法的形式高速、高效进行。

再强调一下，数据的拍卖有两个核心问题需要解决：

- 把数据卖给谁？
- 对每位中标者收多少钱？

对于我们提出的新算法，要解决以上两个问题，首先要将竞标者的标进行一个预处理——按照竞标者提交的最多共享人数对标进行分类。每个标可以出现在不同的类里。例如，图 6.12 所示有 5 个竞标者，每个竞标者下方的第一个数字是他的出价，第二个数字是他能接受的最多中标者（包括他自己在内）。我们按照竞标者允许的中标人数将竞标者分成 5 类，注意，这种情况下没有办法有 4 个或 5 个人同时中标，因为没有 4 个或 5 个竞标者同意 4 个或 5 个人共同中标。

分类过后，我们对每一行进行传统拍卖，英式拍卖、荷兰拍卖、GFP、GSP 或是 VCG，最后综合每行的拍卖结果，最终选择收益最大的一行，作为最终拍卖结果。

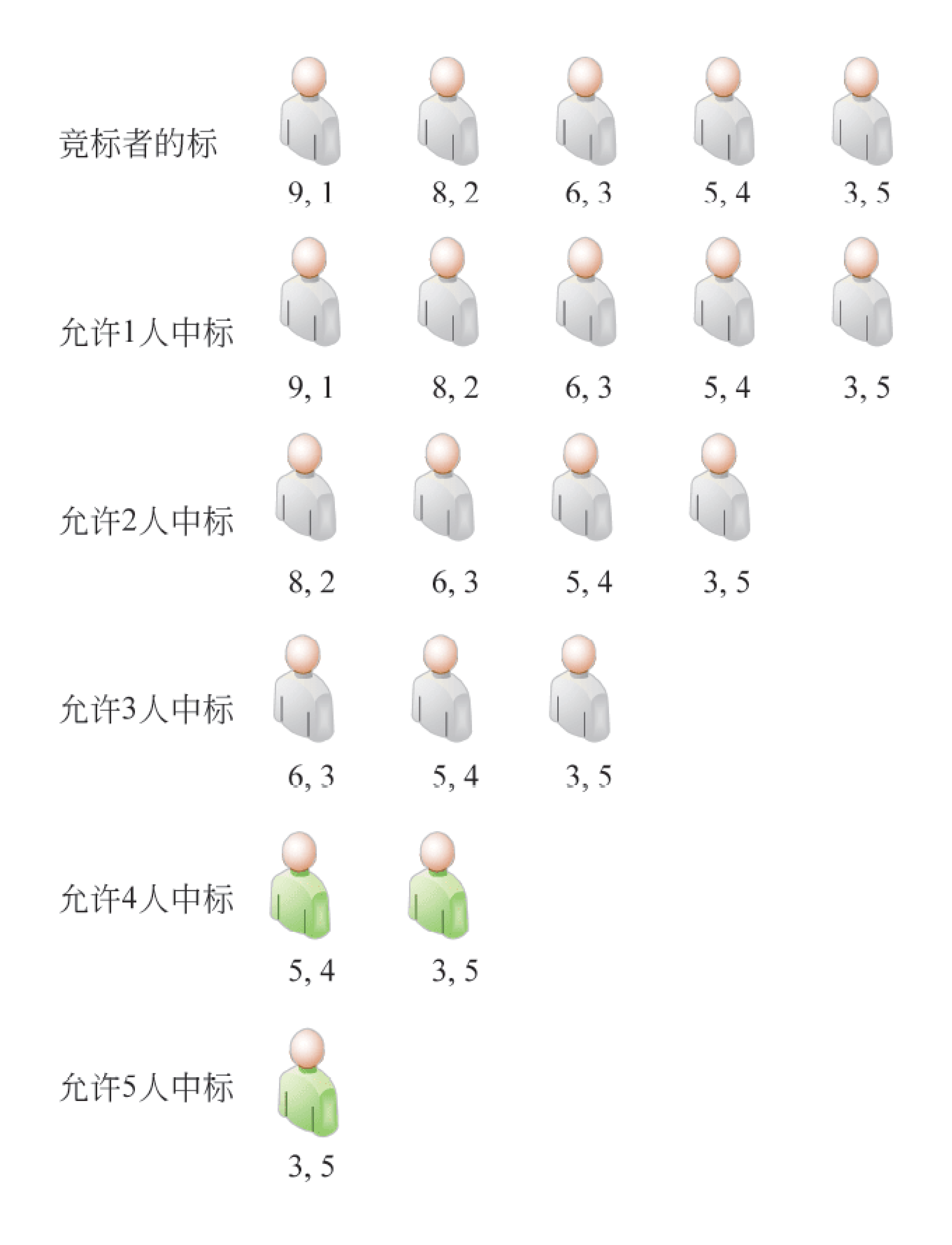


图 6.12　标的预处理（分布式分配拍卖）

假设对每一行做单价拍卖——中标者支付同一价格。当 1 人中标时，拍卖者的收益为 9；当 2 人中标时，拍卖者的收益为 6×2＝12；当 3 人中标时，拍卖者的收益为 3×3＝9。所以最终中标人数为 2 人，分别是出价为(8,2)和(6,3)的竞标者，他们均需支付 6 元，此时拍卖者的收益为 12 元。

假设对每一行做多价拍卖——每个中标者支付的价格不尽相同，这里我们以 GSP 拍卖为例。当 1 人中标时，拍卖者的收益为 8；当 2 人中标时，拍卖者的收益为 6＋5＝11；当 3 人中标时，拍卖者的收益为 5＋3＋1＝9(这里没有比(3,5)低的出价，设定此中标者支付 1 元)。所以最终中标人数为 2 人，分别是出价为(8,2)和(6,3)的竞标者，他们分别需支付 6 元和 5 元，此时拍

卖者的收益为 11 元。

当然,以上只是两个简单的算法。在真正的拍卖中,有多种多样的算法供我们选择。比如刚刚我们求解的过程都是确定的,其实还可以在整个过程中添加随机的元素——在预处理之前,对标进行一次筛选,每一个标都和一个参考价格比对,如果标价大于参考价格就留下,否则直接宣布其失败;参考价格的计算过程是随机的,对于每一个标,都先随机选取除此标以外的一定数量的标,用它们的出价计算出一个参考价格;这样的拍卖便加入了随机的因素。实验证明,对于我们提出的这种新的拍卖,随机拍卖和确定拍卖收益相差不多,并且都较为稳定。所以我们可以根据真实场景中拍卖系统的复杂程度和投标人数来设定拍卖算法。

数据还有一个特性,它并不是一次性售卖商品,它可以重复多次售卖,甚至是持续售卖。所以,除了上述考虑外,每隔多久进行一次数据包或应用编程接口的拍卖也是个问题。如果以数据包的形式拍卖,那么数据与其他数字商品的差别在于时效性,对于有时效性的数据,随着时间的推移,数据会贬值;如若以应用编程接口的形式售卖,数据又不同于其他数字商品,此时,每次拍卖的中标者获得了访问相同数据集的权限,但由于数据是动态产生的,各个得标者通过这个接口获得的数据是不同的。需求方首先需要在竞标前预估需要接口权限的时间,并将此估值作为标的一部分提交,得标后,根据使用情况,需求方可决定是否参与下一次竞标,并调整标中各个信息(出价、使用时间、可共享人数)的值。对于以应用编程接口形式售卖的数据,做成类似计算广告一样的交易系统是完全可以实现的。而对这个系统的建立,最重要的便是机制和规则的设立,保证系统的中立性,实现交易市场的均衡。

从"人的博弈"走向"系统的运行"

"我不害怕计算机，我害怕没有计算机。"

——艾萨克·阿西莫夫，作家

人脑的容量没有那么大，没关系，交给计算机来记；人脑的运算速度没有那么快，没关系，交给计算机来算；人不想，也不能 24 小时无休的做一件事，没关系，计算机可以。是的，只要人类想出了办法，剩下的工作交给计算机就可以了。所有规则确定、逻辑清楚的事情都可以变成算法，由计算机替你执行。几个程序员连夜奋战之后，这件事就变成了只需要按个开关，敲几下键盘就可以完成的了。人类消耗大量脑细胞或者钢笔水的博弈，就变成系统的自动运行了。一个真实的例子就是互联网广告。广告是数据价值商业化最典型、最成熟的应用。从传统的口头广告到如今的互联网广告，数据为广告赋予了新的生命。广告已经不再是从前的广告了。从广告的演变和广告系统的发展中，我们也可以窥见数据交易的发展方向。

广告不再是从前那个"广告"了

表面看上去，互联网广告只不过是把广告放到了互联网上，与传统广告相比，只是换了个媒介而已。但本质上，广告却已经不是从前那个广告了。毕竟网页那么多，商家那么多，分散在全世界各地，谁也没空随时约个下午茶，成百上千的人坐在一起聊聊广告的问题；也不会每隔几分钟就召集商家开个拍卖会，为一个网页上的广告位大费周折，拍卖会还没结束，用户早都关

闭网页了。广告竞价算法背后，是一整套计算广告系统。

也许我们都曾遇到过临街叫卖和发传单，它相比于电视、报纸广告的好处在于，发广告的人可以真实地看见用户的样子，有针对性地发。但是，这也有两点问题。一是有许多特征无法根据外貌判断出来的；二是即使有从外貌能判断出来的特征，但特征也不等同于需求。就像不是所有心宽体胖的人都想或都需要减肥，有时可能会被客户视为骚扰。

那么，怎样才能知道客户的真实需求，并将需要的东西呈现在他们面前，同时又不给客户造成尴尬和困扰呢？这就要靠计算广告了。如果客户最近几天用某种浏览器时常登录某种商品的官网，并且曾经进入支付页面，或者通过某个搜索引擎搜索过某类商品的关键字，算法就能够根据这些数据判断出客户的需求。这时，搜狗浏览器、百度搜索就成了广告主投放广告的最佳地点。除此之外，我们还可以通过客户经常使用的软件来决定广告投放媒介。

当然，这些广告不会简单粗暴地直接呈现在每个用户面前，提供广告位的平台商有义务帮助广告主找到准确的客户，并且只向他们推荐这些广告。广告平台需要不断地通过收集和分析每个用户的特点和行为，给每个用户贴上标签，以便广告主决策广告投放的客户群体。计算广告算法令广告精准，每个客户都可以看到自己想要的广告，同时在同一个广告位，也可以在同一时间，向不同客户展示不同的广告。这里面使用的最主要的技术之一就是受众定向。实现这一技术最主要的就是大数据的支持。

上面提到，计算广告的数据主要来自广告主和媒体网站。媒体网站相当于传统广告的广告牌，是提供广告位的网站。假设淘宝网想在百度搜索上投放广告。作为媒体网站的百度搜索，掌握着第三方数据，对于百度的用户，它精准地知道他们此刻搜索了什么内容，最近经常搜索什么内容，从而推断用户的需求；而身为广告主的淘宝网自己有第一手数据，对于自己的用户，淘宝

知道他们是男是女，喜欢什么样的风格，最近想买什么。如果只利用淘宝的数据，那么将不知道在什么时间、给哪些用户投放广告；但如果只利用百度搜索的数据，有些特定需求又没办法满足，比如淘宝想挽回流失的用户，或是希望百度能帮它找到与它本来用户相似的潜在用户。因此，广告主的数据将极大地帮助定制化用户标签的加工，使广告能够更精准地进行投放，并且更加符合广告主的需求。

有了媒体网站和广告主的配合，实时竞价在程序化交易市场中如鱼得水。

实时竞价需要 ADX 和 DSP 两个平台的支撑。ADX（Ad exchange）是广告交易平台，负责接收竞拍者提交的信息，得到广告候选，而后根据算法选出中标者。整个拍卖的过程由 ADX 来做，这样一个系统就省了拍卖场地、省了人力物力、极大地加速了拍卖流程。ADX 是按照 CPM 收费的，即按每千次广告展示收费。DSP（Demand Side Platform）是需求方平台。在实时竞价中，需求方能够完全地表达自己的投放意愿，它的技术和算法比 ADX 要复杂。DSP 先从广告库中检索想要参与竞争的广告，然后对广告进行排序、定价，最后向 ADX 报出候选广告和对应的出价。DSP 按照 CPM 向 ADX 报出广告的出价，所以准确地预估出 CPM 对出价来说是非常重要的。整个拍卖过程中对 CPM 的估计都由 DSP 来承担。但是 DSP 不能盲目地将 CPM 的值作为出价提交给 ADX，它需要根据 CPM 和市价随着时间的变化来决定最终的出价。此外，DSP 结合媒体网站和广告主的用户标签做受众定向，同时也会帮助拉拢老顾客，发掘新顾客。

图 6.13 展现了实时竞价的过程。

用户进入媒体网站，网站向 ADX 发送请求：“放哪条广告呀？”ADX 向各个 DSP 发送数据，问：“你们都出多少钱？”DSP 根据传过来的数据和广告主自己的数据决定要不要参加竞标，如果参与，便把自己的出价发给 ADX。

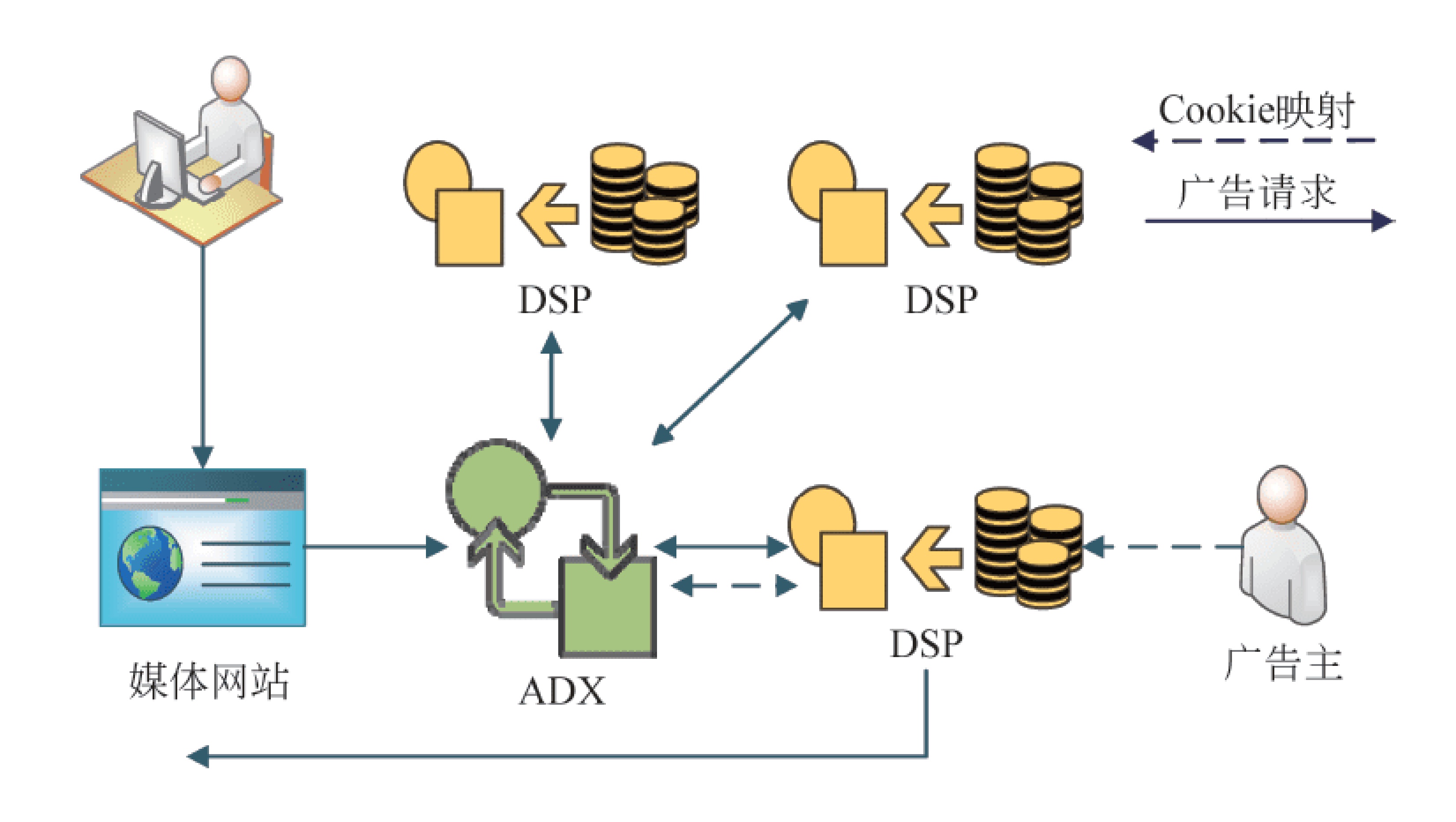


图6.13　实时竞价过程示意图

ADX等待事先规定好的时间，然后按照GSP拍卖选出中标者，并将结果返回给媒体网站。最后，媒体网站从DSP那里拿到广告，在自己的网站上完成这一次展示。

这里要说明的是，当ADX给DSP发送用户标签数据时，DSP如何知道ADX发来的用户数据是对应自己手里的哪些用户的呢？这就要用到Cookie。Cookie是某些网站为了辨别用户身份、进行用户路径跟踪而储存在用户本地终端上的数据。也就是说，每当用户访问一个网站的时候，网站就会给他设置一个用户ID，相当于起一个名字，当用户跳转到网站中其他网页的时候，这个ID就可以用来跟踪他的浏览记录。当然啦，为保护用户隐私并提高安全性，Cookie通常是要经过加密的。我们通过一个例子来看看在计算广告中Cookie是怎么发挥作用的。

假设韩梅梅的计算机里原本没有任何Cookie记录。她访问了www.baidu.com的首页。百度首先会确认此时此刻这个广告位是可以动态分配的，然后ADX会进行发送消息给各个DSP，告诉他们可以来竞争广告位了。经过一系列的计算和厮杀后，有一个DSP赢得这次拍卖，并将自己的广告回复给ADX。ADX向韩梅梅展示此DSP的广告，并在系统中设置她的

Cookie。浏览器调用百度的 Cookie 映射服务，读取韩梅梅计算机中的 Cookie，并将加密后的用户 ID 发给此 DSP 设置的 Cookie 映射 URL。这个 DSP 生成 Cookie，并将此 Cookie 存储在其映射表中与韩梅梅的用户 ID 相对应的位置。此 DSP 将其 Cookie 放到韩梅梅的浏览器中。

一个星期后，韩梅梅再次访问了 www.baidu.com。现在，韩梅梅的计算机上同时存有 DSP 和 ADX 的 Cookie，我们来看看它们是如何匹配上的。

韩梅梅会看到网页，同时，网页代码会包含向百度请求广告的调用。在广告竞价期间，ADX 会向实时出价合作伙伴 DSP 发出调用请求，问它是否要出价投放广告。DSP 收到包含展示信息和用户 ID 的广告调用，在其匹配表中查找韩梅梅的用户 ID，以找出一周前创建的 Cookie。然后，这个 DSP 利用与其 Cookie 相关的信息，对广告位进行出价并赢得这次展示机会。DSP 根据所掌握的信息向韩梅梅投放与其兴趣相符的广告。

广告里的数据交易

计算广告系统是最早进行数据交易的平台。

我们来看图 6.14，ADX 在问 DSP 愿意出多少钱买广告位的同时，也会将相应的数据发给 DSP，使得 DSP 在对广告位出价时，能够结合广告主自己的数据和提供广告位的媒体网站的数据，更精准地做决策。理论上，只有最终中标的广告主可以继续使用这份数据。然而，实际上，要保证其他广告主没有偷偷使用数据，还需要完善相应的政策和技术。

在线数据交易系统

我们已经多次提及，数据时代真正有意思的事情是数据变得在线了。

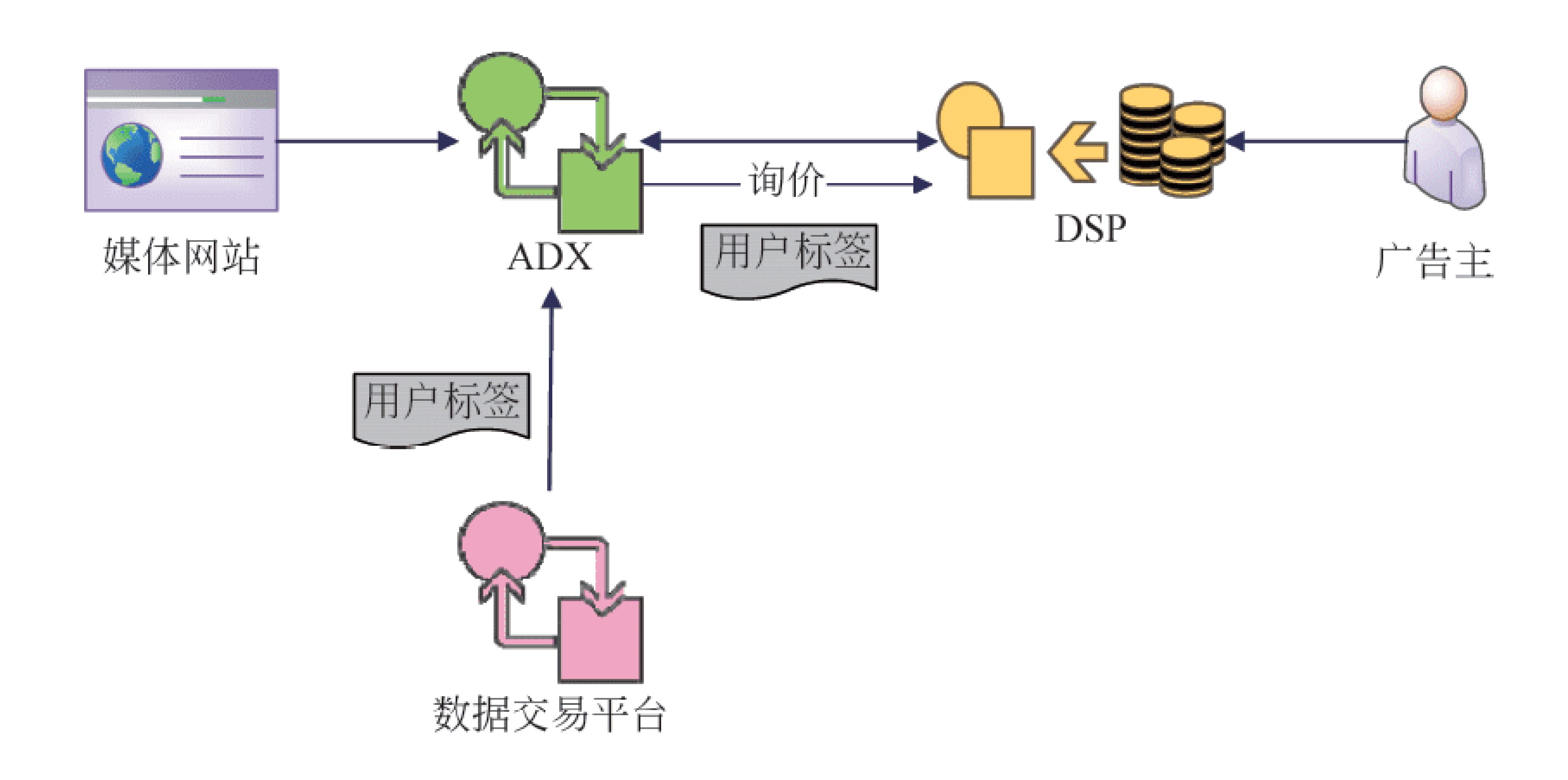


图 6.14　广告系统中的数据交易示意图

虽然数据交易平台不断涌现，数据交易案例也逐渐增多，但是目前数据平台的交易还是以线下为主、线上为辅。显然，这并不是一个高效的方法。而对于数据拍卖，拍卖者和竞标者坐在一个会场里举手喊价这种做法也有种穿越回到古代的感觉。受到计算广告系统的启发，我们尝试让数据通过程序化的系统进行交易，这样我们就可以通过建立一套独立的算法，专门用于数据交易。

大数据的本质是在线。数据交易也应在线，实现程序化、自动化、规模化，这样才能使数据流通更加快速地产生价值。

目前的数据形式一种是数据包，一种是应用编程接口。对于以数据包形式出售的数据交易，交易过程可以参照图 6.15。数据供给方将数据出售请求发送给数据交易平台，说："我有数据要卖。"随后，数据交易平台将供给方提供的数据规格（规模、数量、格式、大小）、数据字段等基本信息和数据样例发送给需求方，同时询问需求方是否购买此数据，出价多少。需求方根据数据基本信息和样例决定是否参与竞争，返回决定，如果参与竞价，则同时返回出价。数据交易平台运行拍卖算法，选择中标者。为了保证数据平台的中立性，在选出中标者后，平台将拍卖结果返回给供给方，然后由供给方直接将数

据交给需求方。

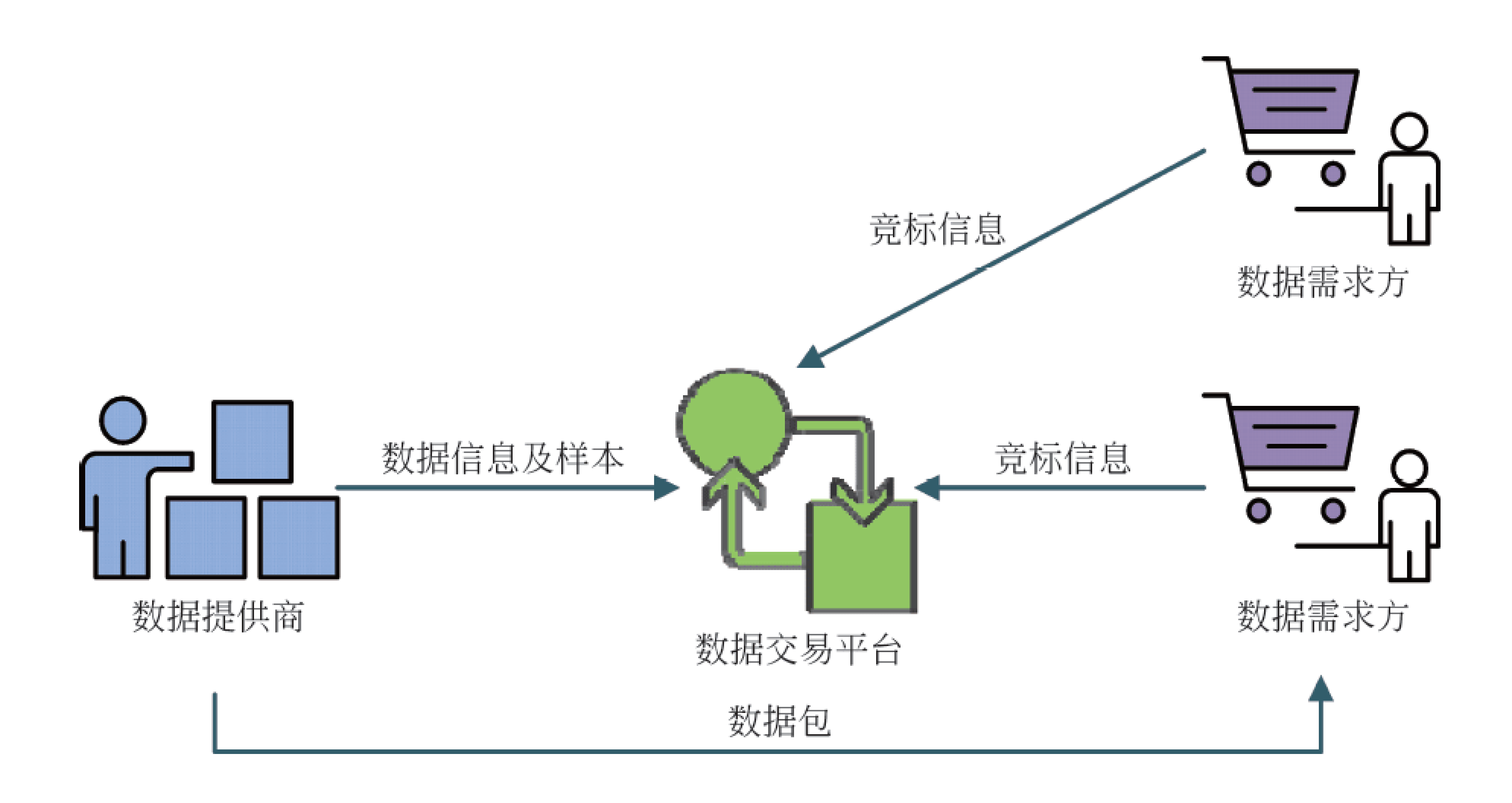


图 6.15　以数据包交易的数据交易系统示意图

对于以应用编程接口形式出售的数据交易(交易过程可以参照图 6.16),数据供给方将访问数据接口的权限提供给需求方。供给方首先确定是以时间为单位出售,还是以抓取次数为单位出售,然后发送请求给交易平台,开启拍卖。交易平台依旧处于中立立场,只负责接收供给方的要求和需求方的出价,运行拍卖算法。由于应用编程接口形式的数据交易频率比较高,并且每次拍卖需要依靠历史交易数据,故可以设立一个数据服务商,以记

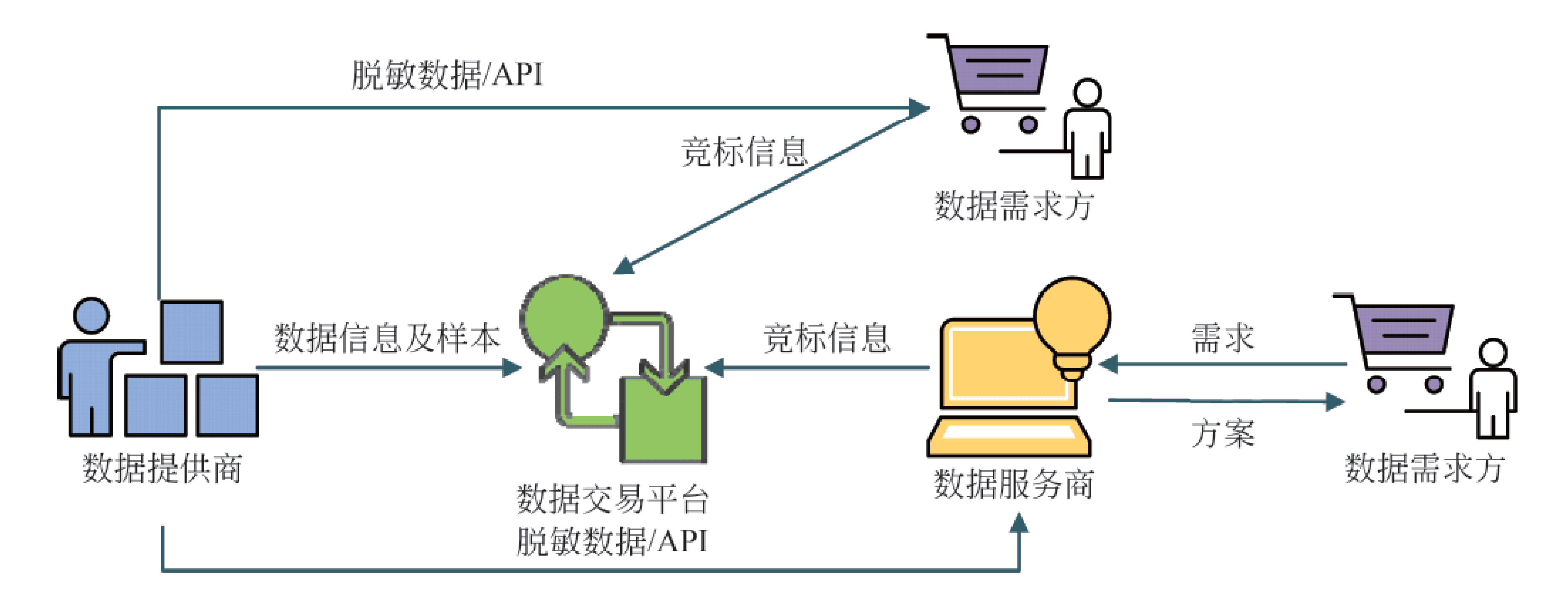


图 6.16　以应用编程接口形式交易的数据交易系统

录历史交易，综合需求方的需求，对数据进行出价。此外，数据服务商还可以兼管数据管理业务，对数据进行统计分析和资产化管理，提供行业解决方案。数据服务商可以是数据需求方的一个部门或分支机构，也可以是单独的承包其数据管理业务的企业。

破解数据交易的不可能三角

政府的当务之急，不是要去做那些人们已经在做的事，无论结果是好一点还是坏一点；而是要去做那些迄今为止还根本不曾为人们付诸行动的事情。

——约翰·梅纳德·凯恩斯

当智能经济从多维空间的科幻电影里走进我们的真实生活时，我们的世界仿佛悄然被偷换了，又好像只是莫名被延展了。当数据在智能经济中大展魔法时，当人工智能、云计算、数据挖掘这些词变成行内行外人都津津乐道的话题时，我们发现，一个真实的智能经济已经在我们的生活里。

数据是智能经济世界的重要支撑。数据像是智能经济世界中的土壤，里面含有多少养分，若不经提取，我们不会知道。对于不同的植物来说，同一块土壤，能给它们的养分也不相同。而养分，就是有物理意义的信息。它不好一概而论地定量，但数据却可以定量。

数据不能被垄断，所有的数据都掌握在一家公司手上，并不是一件利于经济系统健康运行的事情。数据交易让数据变得可以流动，提高了智能经济系统的效益。

数据交易从概念的提出，到平台初见其成，受到了越来越多的人的关注。有人提出了数据交易中的“不可能三角”。

“不可能三角”一词来源于经济学，用来说明美国经济学家保罗·克鲁格曼提出的“三元悖论”原则。当然，也有人说是蒙代尔提出来的。“不可能三角”，是说一个国家不可能同时实现货币政策独立性、汇率稳定以及资本自由流动三大金融目标，而只能同时选择其中的两个。也许你听过某国的总统竞选时提到要同时实现上面三个目标，事实上那大概只能停留在总统候选人美好的政治愿景中。“不可能三角”也反映了世事万物背后对立又统一的矛盾概念，为了得到必须取舍，相生相克，相辅相成。

数据交易也是一样。咨询顾问胡嘉琪指出，在数据交易中也存在这样的三角：严格的数据使用授权、数据的大规模流通以及提升数据的使用价值。

严格的数据使用授权明确规定了数据的所有权和使用权，数据交易必须经过合法授权。在数据交易黑市中，高价值的数据大规模流通，但因泄露用户隐私造成的违法犯罪不在少数。明晰的授权使得用户隐私保护有所凭依，也为数据交易提供了法律上的保障。数据的大规模流通提升了数据价值。但是，要做到同时保障数据授权和数据的使用价值，则会大大增加数据交易的成本，从而对数据规模的发展有所束缚。而在一些实现了大规模流通且相对严格地保障了数据使用授权的情况中，例如政府数据开放平台，数据难以具备广泛的使用价值，再如提供公共数据查询服务的应用编程接口集市，其数据往往欠缺附加价值。

那么，数据交易中的“不可能三角”是无解的吗？笔者认为不然。数据虽然是特殊的商品，但也具有商品的一般属性。要破解上述“不可能三角”，需要从建立良好的数据交易生态圈入手。

对有数据处理、分析能力的提供方和需求方来说，数据质量是有所保障的，数据的使用价值和附加价值也能够被深度挖掘；而对于没有这些能力的

提供方和需求方来说，数据服务商和数据交易平台可以为其解决数据质量判断、价值提取和数据增值等问题。也许有人会说，这样不更加提高了数据交易成本，抑制了数据的大规模流通了吗？这一点大可不必担心。数据的需求量越来越大，在授权明晰的数据交易市场中，合理、有效的数据定价，可以大大降低数据交易成本，最大化数据的价值，从而扩大数据交易规模。例如我们提出的在线数据交易系统，以及可以运行在系统中的新型的拍卖算法，都将大大地提高数据交易效率，降低数据交易的门槛，打造良好的数据交易生态圈。如此，“不可能三角”就不再是数据交易的瓶颈了。

值得关注的是，即使明确了数据使用权，市场仍然要建立保护机制来防止侵权行为。那么，如何用技术手段保证数据的安全并保护用户的隐私呢？我们将在第 7 章继续讨论。

第 7 章　谁来保护我们的隐私

1969 年，在美国加州大学洛杉矶分校，坐在接口信息处理机（Interface Message Processor，IMP）①前的大学生查理·克莱恩（C. Kline）怀着激动不安的心情，在键盘上敲入了一个字母 L，然后对着长途电话另一端、位于斯坦福研究院的终端操作员喊道："你收到'L'了吗？"

"是的，我收到了'L'。"操作员兴奋地回答。

"你收到'O'了吗？"

"是的，我收到了'O'，请再传下一个！"

克莱恩紧接着键入了第三个字母"G"。然而这时传输系统突然崩溃了（图 7.1 是当时的工作日志）。虽然这次互联网的通信试验仅仅只传送了两个字母"L"和"O"，它的意义却非常重大且激动人心，因为这就是我们今天无时无刻不在使用着的互联网的最初起点。

---

① 1969 年，美国 BBN 公司制造了第一台接口信息处理机（IMP），也是阿帕试验网的第一个节点。

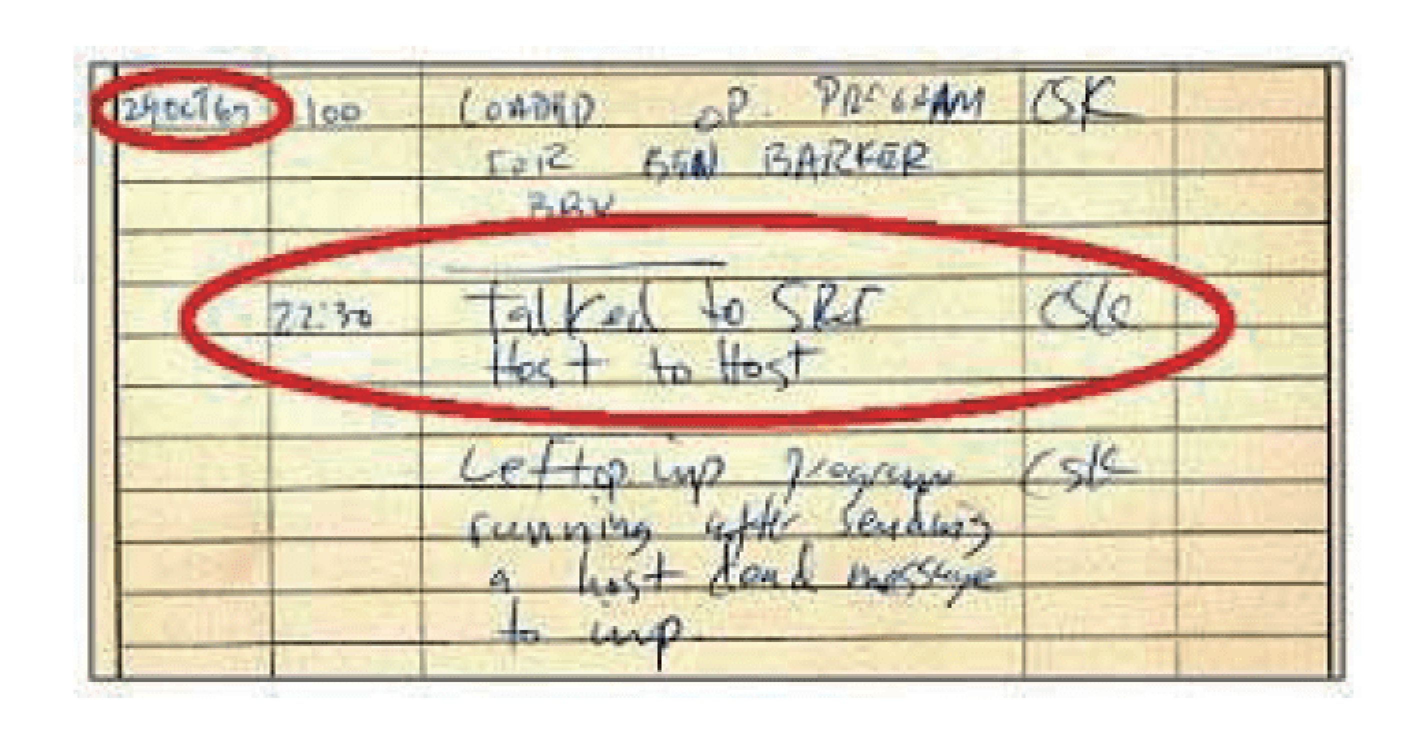


图 7.1　见证首次互联网连接实验的工作日志

最早创建互联网是为了实现信息交换，如今已经过去 40 年了，人类制造出的信息量已经要用 ZB(1ZB＝1024TB)计算，这比全世界所有海滩的沙粒数还多。大数据技术的发展，让这些高速流转的海量数据可以为我们提供所需的服务，我们的日常生活轨迹也越来越详细地被数据记录和计算。然而，当一切变得有据可查，有迹可循，可供推演，这些关乎你的生活和隐私的大数据，会不会也令你感到前所未有的担忧？

我们还有隐私吗

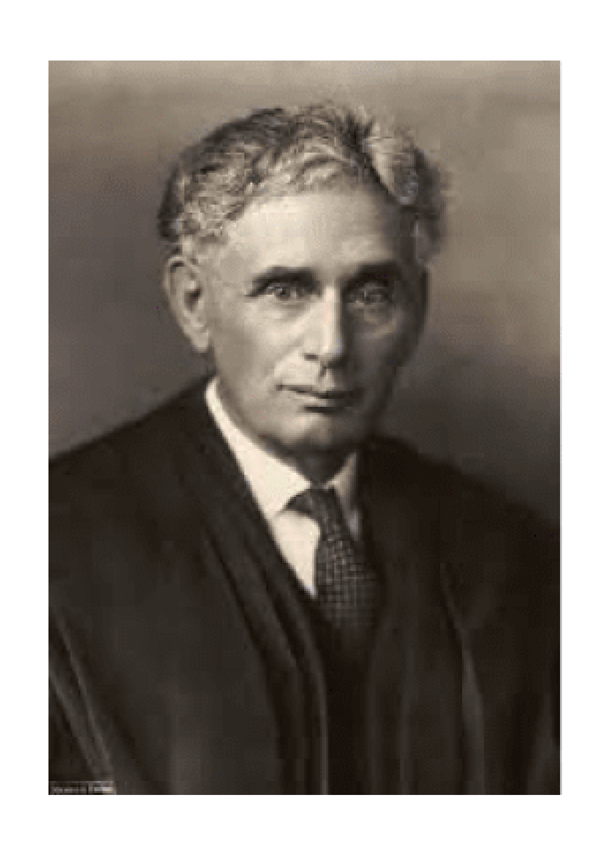
在巴黎，阶沿上有耳朵，门上有嘴巴，窗上有眼睛；最危险的莫过于在大门口讲话。彼此临走说的最后几句，好比信上的附笔，所泄露的秘密对听到的人跟说的人一样危险。

——巴尔扎克

隐私权是基本人权

技术的进步不断侵占人们的隐私，而在保护隐私的问题上，人们一刻也没有停止过斗争。

1890 年 12 月，路易斯·布兰代斯(Louis D. Brandeis)与他的法律合伙人塞缪尔·沃伦(Samuel D. Warren)在《哈佛法律评论》上发表了《隐私权》一文(图 7.2)。这是美国历史上最为著名的法学论文之一，也是隐私权在世界上首次被提出。截至 2017 年 7 月，这篇论文的 Google Scholar 引用数量已经超过 9700 次。





HARVARD

LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make common law without a precedent; much more when received and approved by usage."

WILLES, J., in Millar v. Taylor, 4 Burr. 2303, 2312.

THAT the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection. Political, social, and economic changes entail the recognition of new rights, and the common law, in its eternal youth, grows to meet the demands of society. Thus, in very early times, the law gave a remedy only

图 7.2 路易斯·布兰代斯（左），塞缪尔·沃伦（中）和他们发表的《隐私权》

美国著名社会工作者弗洛伦斯·凯利曾经说过：“在林肯之后，再没有人比路易斯·布兰代斯更理解人民大众了。”布兰代斯曾是美国最早提供无偿公益服务的律师之一，凭借丰富的社会科学知识和对现代工商业运作和规范的了如指掌，他极力论证过最长工作时间和最低工资的合宪性，他坚持不懈地与那个时代的托拉斯、垄断和其他强悍的商业利益既得群体做斗争，被视为“人民自由的捍卫者”。1916 年，美国总统威尔逊任命他为美国最高法

院的终身大法官，他也成为美国联邦最高法院历史上第一位犹太裔大法官。

布兰代斯的父母来自东欧，早年因为奥地利一场反犹太主义运动移民美国，布兰代斯出生在美国肯塔基州的路易斯维尔城，18 岁时，他进入美国哈佛大学法学院与沃伦成为同学。他们共同协助创办《哈佛法律评论》[①]（见图 7.3），1877 年毕业后，他们合伙开了一家律师事务所。

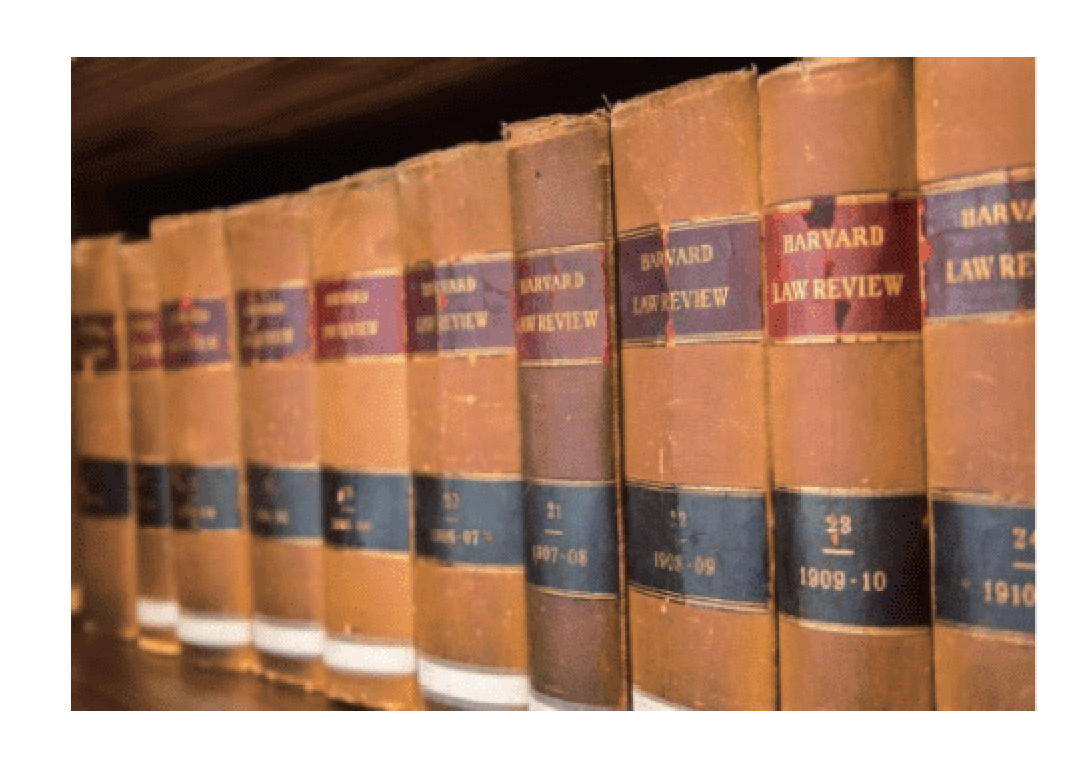




图 7.3　《哈佛法律评论》的历年藏本

当时的美国，黄色新闻猖狂，为满足一般人对上流社会生活的好奇，报纸大量刊出名人私生活的相关消息迎合大众口味。其中，沃伦的夫人梅布尔·贝尔德（Mabel Bayard）就是被关注的对象之一。沃伦夫人出身名门，她的父亲是一位参议员，她时常在家里举办聚会、派对等社交娱乐活动。波士顿报纸热衷于对沃伦家里派对的具体细节情况进行详细报道，一并掺杂着一些男女之事的描述。自 1882 至 1890 年间，共有近 60 篇关于沃伦·贝尔德家族生活的私密消息被披露给公众，特别是沃伦的母亲和姐姐时隔两周的葬礼也上了报纸头条。为此，沃伦异常愤怒：他的家人受到了侵犯，他的家族隐私就像是时刻暴露在长焦镜头之下。他邀请布兰代斯共同撰写并发表了《隐私

① 1887 年，哈佛大学法学院一位名为 J. McKelvey 的 3 年级学生发起了《哈佛法律评论》——美国第一家由学生编辑和管理的法学刊物，作为美国法学最高研究水平的代表，美国法官常常引用该刊的评论文章论证判决的正当性。

权》文章，其中写道："新闻报刊超出了礼义廉耻可以容忍的限度。传播流言蜚语不再是闲散无聊人士的消遣，而成为一种行业，被人们孜孜不倦又厚颜无耻地从事着。"文章主要观点[①]有：

不论表达形式是言语、符号、绘画、雕塑或是音乐都无关紧要。这一权利的存在也不取决于思想或情感的性质与价值，或表达手段有多么杰出。一封寻常信件或一篇日记记录，与最有价值的诗篇或散文；拙劣的拼凑或涂鸦乱抹与名著杰作，它们所获得的保护别无二致。在前述的每一种情况之下，**个人有权决定是否将属于自己的东西公之于众**。

如果说这些判决阐明了思想、情绪、感情上的一般性隐私权利，那么，无论其表达方式是写作、举止、交谈、姿势或面部表情，都应当受到同样的保护。**隐私权，是更为一般的个人受保护权(即人格权)的一部分**。

保护个人作品以及其他智力产品、情感产品的法则，是隐私权；法律无须阐明新的原则就可以将保护范围拓展至仪表、言语、行为以及和家庭及其他领域的个人关系。

如果侵犯隐私权构成法定的侵权行为，从侵权行为本身所致的精神痛苦的价值被确认为赔偿的原因之后，要求损害赔偿的基础便得以存在。

沃伦和布兰代斯在文中论述和援引了许多普通法的判例，推导出隐私权的存在，他们认为古老的普通法在永恒的青春中不断成长，随着历史的演进，政治、社会和经济的变化促使新的权利"隐私权"产生并最终成型于现代社会。文明的前行带来了日渐紧张和复杂的生活状态，人们对公共场合变得更加敏感和有界限感，在此基础上，拥有一个自己的相对独立和私密的空间显得越来越重要。但是，新近的发明和商业手段一定程度上为侵犯人们的隐私

---

① 这些观点主要参考 Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, Harvard Law Review, Vol. 4, No. 5, Winter 1890。

提供了便利的工具和途径，隐私被侵犯使人们遭受精神上的痛苦与困扰，与以往纯粹身体上的伤害相比，有过之而无不及。因此，人们迫切地需要进一步采取措施来保障个人“不受打扰”的权利。

第一部隐私权法案诞生后，作为满足社会进步和维护公民利益的基本权益，隐私权保护不断向前发展。1960 年，威廉 · 普雷瑟（William L. Prosser）开创了隐私权的分立理论，美国侵权法中隐私权基本体系建立。1965 年，格鲁斯沃德（Griswold V. Connecticut）诉康涅狄格州案中，道格拉斯大法官（Douglas）对法院提议，形成了宪法层次个人隐私权的存在，正式宣告隐私权是受美国宪法保护的一项基本权利。1974 年美国国会制定了《联邦隐私权法》，1986 年通过了《联邦电子通信隐私法》，2000 年出台了第一部关于网络隐私权保护的重要法律《儿童网上隐私权保护法》，还有一些相关法规包括《联邦记录法》《金融隐私权法》《视频隐私法》《电话消费者保护法》《驾驶员隐私保护法》《电信法》等，希望通过采用政策引导的行业自律模式保护公民网上的个人信息隐私权。

2013 年 12 月 18 日，联合国大会通过了一项“数字时代隐私权”的决议。该决议强调，隐私权是民主社会的基础之一，非法或任意监控通信以及收集个人数据，是侵犯隐私权和言论自由权利的行为，背离了民主社会的信念。该决议还要求各国建立有效的国内监督机制，确保涉及通信监控和截取，以及对个人数据收集的透明度，并接受问责。

这项决议虽然没有强制效力，却在政治和道德层面体现了国际社会对保护网络和电子通信使用者隐私权的态度，表达了对越界情报行动以及对个人数据大规模搜集的批评与忧虑。

2016 年 4 月 14 日，在法国的斯特拉斯堡，欧洲议会正式通过了商讨四年之久的《一般数据保护条例》，它取代了 1995 年发布的《欧洲数据保护指令》，并直接适用于欧盟各成员国，意味着欧盟对个人信息保护及其监管达到

了前所未有的高度。

新条例按照数据的分布来认定法律管辖范围,与提供服务的企业所在国家和地区无关,即法律管辖范围会适应数据流动的特性而变化。同时,新条例还提出了“被遗忘权”,明确了数据所有者如果不希望自己的数据由互联网企业使用,可以撤回数据的使用权,并要求完全删除相关数据。

2014 年 10 月 10 日,我国开始实施最高法院《关于审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定》,这是首次以法律的形式为我国公民的个人信息划定了“保护圈”。2016 年 11 月 7 日,十二届全国人大常委会表决通过了《中华人民共和国网络安全法》(以下简称《网络安全法》)。对于个人隐私信息保护,《网络安全法》制定了多个条目,包括“网络运营者收集、使用个人信息,应当遵循合法、正当、必要的原则,公开收集、使用规则,明示收集、使用信息的目的、方式和范围,并经被收集者同意。”“网络运营者不得收集与其提供的服务无关的个人信息,不得违反法律、行政法规的规定和双方的约定收集、使用个人信息,并应当依照法律、行政法规的规定和与用户的约定,处理其保存的个人信息。”“网络运营者不得泄露、篡改、毁损其收集的个人信息;未经被收集者同意,不得向他人提供个人信息。”

在全球范围内,隐私权作为一项基本人权,受到越来越多的关注和保护。

当你的一切都可能被泄露

美国对隐私的关注已经有一百多年,然而,它作为最早提出隐私权概念的国家,却也是臭名昭著的“棱镜”计划诞生的地方。**那些沃伦和布兰代斯时代完全无法想象的科技,正在为获悉个人最为私密、最为个性的信息提供更多可能性,现在比以往任何时候都要猖獗。**

2016 年下半年,希拉里作为美国历史上第一位主流政党的女性总统候

选人，与特朗普共同角逐总统竞选。参选初期，希拉里的支持率一路领先，曾获得 13％竞选优势，然而正当她节节胜利之时，维基解密的好事者截获并分析了她的私人邮件，这些通过私人邮箱和位于家中的私人服务器收发的 3 万封邮件，包括了许多涉及国家机密的绝密邮件，虽然美国联邦调查局 FBI 最后并未提出指控，然而维基解密的创始人阿桑奇①对邮件门的追查并没终止。这位愤青技术宅人恢复出部分被删除的“私人邮件”，同时将用技术手段获取的其他 7 位民主党的重要官员在内的约 2 万封内部邮件一起在维基解密上予以公布。被爆料的邮件显示：希拉里曾不遗余力地搞垮她的党内竞争对手，有计划地、有系统地伪造特朗普的负面新闻；还与其团队私下从事着明码标价陪吃饭和买官卖官的活动，如图 7.4 和图 7.5 所示。

Sum of Amount	Column Labels				
Row Labels	2013	2014	2015	2016	Grand Total
Breakfast	$ 231,203.00	$ 4,950.00	$ 250.00		$ 236,403.00
Brunch	$ 645,203.00				$ 645,203.00
CANCELLED	$ 404,785.00	$ 257,604.44	$ 250.00	$ 250.00	$ 662,889.44
Clutch	$ 127,400.00	$ 87,400.00			$ 214,800.00
Concert	$ 355,150.00			$ 4,450.00	$ 359,600.00
Conference	$ 673,651.00	$ 653,305.00	$ 439,005.00		$ 1,765,961.00
Dinner	$ 2,409,560.00	$ 3,521,839.20	$ 3,042,866.00	$ 1,486,050.00	$ 10,460,315.20
Discussion	$ 5,783,372.37	$ 12,417,148.00	$ 7,100,699.90	$ 767,600.00	$ 26,068,820.27
Gala	$ 684,955.00	$ 910,743.00	$ 1,435,249.33		$ 3,030,947.33
General	$ 1,779,687.74	$ 2,214,848.60	$ 3,824,918.99	$ 3,863,156.00	$ 11,682,611.33
In-Kind	$ 6,000.00			$ 225.00	$ 6,225.00
Lunch	$ 1,585,438.40	$ 11,615.00	$ 75,850.00		$ 1,672,903.40
Reception	$ 4,527,360.00	$ 5,695,117.78	$ 3,150,197.18	$ 2,96[illegible]0.75	$ 16,338,555.71
(blank)			$ 3,010,015.50	$ 14,328,[illegible]	$ 17,338,621.93
Grand Total	$ 19,213,765.51	$ 25,774,571.02	$ 22,079,301.90	$ 23,416,218.18	$ 90,483,856.61

图 7.4　民主党靠饭局在历年获得的收入

面对如此劲爆的消息，美国却没有一家主流媒体进行报道，美国民众看不到希拉里的丑闻，就连一向表明独善其身的谷歌公司也明显偏向了希拉里，例如，当时在谷歌搜索引擎中输入“总统候选人”竟然都不会出现特朗普。更加离奇的是，美国民主党的一名数据主管人员康拉德 · 里奇（Seth Conrad

① 阿桑奇，16 岁就潜入加拿大电信系统的天才黑客，后创立维基解密网站。

	Donations as of 11/28/2008				
	Name	Total Raised	Position	Date Start	Date End
1	Matthew Barzun	3,503,080	Ambassador UK / Ambassador Sweden	8/21/2009	Present
2	Julius Genachowski	3,494,919	Chairman FCC	6/29/2009	11/4/2013
3	Frank Sanchez	3,415,000	Under-secretary Commerce	3/29/2010	Nov-13
4	Jeffrey Katzenberg	3,120,500			
5	Frank White	2,983,005			
6	Stanley Grinstein	2,592,000			
7	Charlie Rivkin	2,562,400	Ambassador France / Asst Sec State	8/3/2009	Present
8	Kirk Wager	2,330,000	Ambassador Singapore	9/25/2013	Present
9	Alan Solomont	2,328,500	Ambassador Spain	12/24/2009	8/1/2013
10	Mark Gorenberg	2,043,075			
11	John Roos	2,030,150	Ambassador Japan	8/20/2009	8/12/2013
12	Nicole Avant	2,008,550	Ambassador Bahamas	10/22/2009	11/21/2011
13	Eleen Chamberlain Donahoe	2,000,625	Ambassador UN	2010	2013
14	Jim Crown	2,000,233			
15	Steve Spinner	1,876,435			
16	Steve Westly	1,814,849	CFO California	1/6/2003	1/8/2007
17	Don Beyer	1,733,169	Ambassador Switzerland	9/8/2009	Present
18	John Rogers	1,685,000			
19	Orin Kramer	1,681,300			
20	Michael Adler	1,565,500			
21	Don Gips	1,547,100	Ambassador South Africa	7/31/2009	1/2/2013
22	Howard Gutman	1,541,050	Ambassador Belgium		
23	Robert Wolf	1,524,450			
24	Cynthia Stroum	1,481,487	Ambassador Luxembourg	12/7/2009	1/31/2011
25	Mitchell Berger	1,462,300			
26	Ayelat Waldman	1,449,643			
27	Mark Gilbert	1,429,750	Ambassador New Zealaand	5/12/2015	Present

图 7.5　美国各种明码标价的官位，几十万美元可以到美国驻中国或俄罗斯的大使馆任职

Rich）在华盛顿自己寓所附近被枪杀，当地警方称这是一起持枪抢劫杀人案件，就此结案……但里奇的随身财物、钱包、手机并没有被拿走。

维基解密继续公开希拉里团队的内部资料——希拉里选票造假。具体操作的方法就是拉上一车非法移民去某个投票站点免费一日游，此外，还存在大量的死人投票，例如某二战老兵去年已经去世，结果还进行了投票。据不完全统计，这届选举有 180 万死人票，还有 280 万人在不同州的重复投票。

阿桑奇一次次在维基解密上揭露美国选举黑幕，有一天他连续发送了多条带有错别字却没有什么实质含义的短句，随后删除了这些短句，遍布全球的阿桑奇粉丝们刨根问底，很快就发现所有的错别字连起来形成了一句求救短语："HELP HIM!"。原来就在总统选举进行到最关键阶段时，重装武器的"警察"包围了阿桑奇所在的厄瓜多尔驻英国大使馆，一直对阿桑奇表示支持的厄瓜多尔政府掐断了阿桑奇的网线，阻止他对外发声。

阿桑奇的粉丝们被彻底激怒了，他们联合起来用自己的方式进行示

威——2016 年 10 月 22 日，在黑客的攻击下，全美有近半数网站直接瘫痪，包括 Twitter、Facebook 等，图 7.6 是阿桑奇的粉丝在示威。希拉里的竞选经理约翰·波德斯塔(John Podesta)曾因点开了一封类似 Google 发给他的警告邮件，而将邮箱密码泄露给了攻击者，此时波德斯塔的邮件也在维基解密上进行了新一轮曝光，美国人民这时发现：美国政府一直声称在寻找当今世界上最大的邪教、活跃在伊拉克和叙利亚的恐怖组织 ISIS 的幕后资金来源，原来，希拉里很早就知道是卡塔尔和沙特阿拉伯政府一直在秘密资助 ISIS，不仅如此，希拉里和她的团队还直接接受了卡塔尔和沙特阿拉伯政府的国外资金。





图 7.6　阿桑奇粉丝们聚集在大使馆门口抗议，蓝色标语："释放阿桑奇"

其中一封邮件显示，卡塔尔承诺将向克林顿基金会"捐助"100 万，沙特阿拉伯在克林顿基金会成立之后，先后"捐助"了 1000 万至 2500 万美元，据报道，这次希拉里竞选，20%的竞选资金来源于沙特阿拉伯。

这批邮件公布后，美国民众彻底坐不住了。2016 年 10 月 29 日，FBI 重新启动对"邮件门"事件的调查，希拉里的支持率开始下跌，最终败选，特朗普赢得了一次不可思议的胜利。

作为一名普通的吃瓜群众，也许我们会认为这些拦截邮件、挖掘隐私的事情一定不会发生在自己这样的小人物头上。然而，网络世界真如我们想象

的这样友好吗？说实话，**即使没有黑客，没有视我们为羔羊的潜在恶意者为我们构造虚假链接和网络陷阱，这些每天看似平淡无奇的网络生活，也早已将我们的一切出卖**。

1998 年，拉里·佩奇(Larry Page)和谢尔盖·布林(Sergey Brin)在斯坦福大学的学生宿舍内共同开发了全新的在线搜索引擎——谷歌。到如今，谷歌已经成为公认的全球规模最大的搜索引擎，它支持使用多种语言查找知识、地图、要闻和股价，可以帮助搜索数十亿计的图片并详读全球最大的互联网新闻组(Usenet)信息存档——超过十亿条帖子，发布日期可以追溯到1981 年。用户几乎可以在瞬间得到相关的搜索结果，请注意，这个时候全球网民数量已经增加到 30 亿。如何获取更多信息，尤其是更多符合当下情景的信息并利用这些信息度过每一天，这样的用户需求早就远远超出了传统搜索查询的范畴。谷歌在不断发展中又推出了不少了不起的产品，其中包括一些“免费”产品如邮箱、地图、云盘和安卓系统等。用户穿梭于谷歌的各项简单易用的免费服务，逐渐习惯了使用这些“免费”产品来安排一天的生活、来往于各地并与彼此保持联系。其实，所有这些“免费”产品的背后，都在引导使用者不断透露关于自己生活轨迹的资料信息。

从我们使用 Google 进行网络搜索开始，就向 Google 敞开了我们的隐私之门。搜索引擎利用我们的信息，可以推出弹窗等形式的广告。2016 年 2 月，Google 对搜索广告进行了公司诞生以来最大规模的改版，将原本呈现在搜索结果页面右侧的广告展示栏左移置顶，在搜索结果前插入不超过 4 条的文字链接广告(之前上限为 3 条)，也就是说，用户只能在屏幕更下侧的位置看到自己想要的信息结果。话说百度很早就在左侧置顶位置投放广告信息了。

据 Bunnyfoot 的研究显示，有 40%的互联网用户无法区分 Google 的自然搜索结果和付费广告。搜索引擎凭借天然拥有的庞大流量，从投放广告上

盈利早已不是秘密，谷歌母公司 Alphabet 2016 年第四季度财报显示，其广告营收增长超过 17%，达到 224 亿美元，占公司总营收的比例超过 85%，广告点击量增长超过 20%，且仍在逐年增加。图 7.7 给出了对用户搜索行为进行眼球移动轨迹追踪所绘制的热力图，可以看出，广告位置是用户关注的热点位置。



图 7.7　对用户搜索行为进行眼球移动轨迹追踪所绘制的热力图

2004 年，Google 推出了 Gmail，为我们提供电子邮件服务，Gmail 以大容量、高稳定性迅速赢得了用户的赞誉，Gmail 号称用户永远不用删除邮件，当前用户的典型邮箱大小已经达到了 15G，Gmail 也是当前为数不多的可以在中国正常使用的 Google 业务。但与此同时，Gmail 也通过扫描并使用自动化方式阅读我们的邮件以获得更多的信息。Google 为什么关心我们的邮件内容？其实就是为了通过分析我们的邮件，获得我们的偏好信息，从而可以

更有效地向我们推送广告，最终获得更大的广告收益。

2014 年 4 月，Google 更新了其隐私政策，增加了下列内容：“我们的自动化系统将分析你的内容(含电子邮件)以便为你提供相关产品功能，比如个性化的搜索结果、量身定制的广告以及垃圾邮件和恶意邮件的检测功能。这些分析会在内容被发送、接收和储存的时候进行。”

一家专门发布网站流量世界排名的公司 Alexa 最新提供了谷歌网站下不同子域名所占的流量份额数据，Gmail 使用的子域名为 mail. google. com，统计数据显示其约占 Google. com 流量的 44. 14%，在谷歌公司内部排序接近榜首(图 7. 8)。由此看出，电子邮件业务对目前的互联网公司仍是至关重要的，别忘了它们可以从无数电子邮件用户和邮件流量中获得隐形收益，例如，它们能将用户导入其他内部服务的平台(Google 曾试图向 Gmail 用户介绍 Buzz 这项产品)。然而，Google 表示是“用户自愿将信息交给第三方供应商，不应再要求对自己隐私进行保护。”也就是说，谷歌认为近 4 亿 Gmail 用户的隐私没有法律意义上的保护。

Where Visitors Go on Google.com

Subdomain	Percent of Site Traffic
google.com	66.32%
mail.google.com	23.33%
docs.google.com	1.45%
adwords.google.com	1.37%
maps.google.com	0.96%

Where do visitors go on google.com?

Subdomain	Percent of Visitors
google.com	60.55%
mail.google.com	44.14%
accounts.google.com	26.36%
docs.google.com	16.75%
drive.google.com	12.12%

图 7.8　2011 年 5 月 (左)和 2017 年 2 月 (右)邮件在谷歌网站中的流量份额占比，数据来自 Alexa

接下来，谷歌开始支持在线储存联络人信息，以此来评估我们的社交关系。想知道我们会打电话给谁，于是就有了 Google Voice，不仅可以追踪我们的每次通话，还能将语音信息转换成文字。为了让人们更加轻松便利地整理照片和视频，谷歌发布了 Google Photos，不到一年的时间，Google Photos

每月的活跃用户就超过了 1 亿,使用者可以在任何设备上随时找到想要的图片,例如,在 Photos 中找到和“拥抱”有关的照片。此外,还有 Google 地图,基于 GPS 导航信息,Google 地图能够追踪我们去过的一切地点。当我们询问地址时,它不只是回答某地点到另一地点怎么走,根据不同的环境和情景,还会给出躲避拥堵的最佳时间段,我们想去的店铺是否开门,以及初次游览的目的地有什么精选推荐。作为终极大杀器,Google 还设计了安卓操作系统,Google 坚信智能手机是时时刻刻带在用户身边的,有了安卓系统,Google 就能轻松地追踪我们一整天的行为。

Google 提醒用户们上传到谷歌的所有内容都将被一视同仁地对待,其对新版隐私政策的阐述如下:

“当你上传、提交、储存、发送或接收内容至我们的服务时,你就赋予了谷歌(以及我们的合作伙伴)在全球范围使用、托管、储存、再生产、修改、创建衍生内容(比如根据你的内容翻译、改编或进行其他修改而得到结果,我们那样做的目的是让你的内容能够更好地与我们的服务兼容)、传播、发布、公开演示、公开显示和散布这些内容的权利。”

不论在 Youtube、Google+还是 Gmail 等服务上,新版隐私政策的标识都已经十分显眼。然而,Google 对用户数据的使用还在继续深入,除了这些服务外,还打算一口气将旗下 60 多种在线服务全都整合形成单一的数据库。Google 执行董事施密特曾在 2009 年接受 CNBC 采访时提及:“在美国境内,我们受到《爱国者法案》的约束,将这些信息透漏给美国官方并不是没有可能。”这也意味着我们所处的网络环境中,除了互联网公司会搜集用户的数据信息,政府也在无时无刻想要监控我们的数据信息。

2013 年 6 月 5 日,美国中央情报局前技术雇员爱德华·斯诺登向英国《卫报》和美国《华盛顿邮报》爆料,揭露了美国国家安全局(NSA)等美国政府部门监视公民隐私的“棱镜”计划项目,斯诺登称该项目自 2007 年启动,美国

国家安全局已经“直接监视美国 9 家最主要网络公司的中央服务器”，一向标榜“坚决捍卫公民隐私和人权”的美国政府，陷入到巨大的舆论旋窝之中。此后一连串的新闻报道更加印证了：全球的网络和数据通信，正处在以美国政府为主导的监控之下。

在英国《卫报》公布的一段长达 12 分钟的专访视频中，斯诺登说道：“你什么错都没有，但你却可能成为怀疑的对象，也许只是因为一次拨错了的电话，他们就可以用这个项目仔细调查你过去的所有决定，审查所有跟你交谈过的朋友，并在这一基础上对你发起攻击，他们就这样怀疑一个无辜的人，把你当成一个做过错事的人来对待。”

斯诺登注定将成为“国家利益”和“公民隐私”之间矛盾演变最具冲击力的一个传奇。多数网民认为他是“英雄”，他勇敢地向大众公开了真相，美国政府则称他“背叛祖国”，是受到一系列“严重刑事指控”的通缉犯。

也许很多人对“国家监控”一事多少有所耳闻，然而当斯诺登真真切切地证实了它的存在时，人们还是难免会不安、排斥甚至抗议。事实上，从古至今，无论哪朝哪代的政府都或多或少采取过监控公民的措施。而这些监控行为一旦曝光，大都会引发“政务机密”和“公民隐私”之间关系的广泛讨论。美国人不由得拿自己与小说《1984》里的主人公温斯顿进行对比，美国奥巴马政府和反恐部门则忙不迭地四处解释，监控系统是为了确保公共安全，并非要偷窥本国民众的隐私。

事实上，保障公共安全确实是事实。美国联邦调查局副局长肖恩 · 乔伊斯在一次听证会上列举了 4 起依靠监听被阻止的恐怖袭击案件，其中包括 2009 年未遂的纽约地铁炸弹袭击。美国国家安全局局长、美军网络司令部司令基思 · 亚历山大也表示：“自‘9 · 11’事件发生以来，通过这些监视项目，我们已成功阻止了近 50 起可能的恐怖袭击，但具体细节就不方便公开了，因为如果我把这些说出去，恐怖分子就会知道我们是怎么追踪他们的。”

然而，CNN 公共安全分析师、新美国基金会国家安全研究计划负责人彼得·伯根（Peter Bergen）反驳称，新美国基金调查了自“9·11”以来有关圣战组织恐怖分子的法庭记录和媒体报道，发现在挫败恐怖袭击的过程中，发挥最主要作用的仍然是传统的合法手段。也就是说，NSA 的监听项目是缺乏说服力的大范围“钓鱼”项目。所以在质疑者看来，仅仅公布四个案例是不够的，公众需要更多的例证来判断他们所做出的牺牲是否值得。

斯诺登踢爆“棱镜门”之后，其他国家的一些监控项目也逐渐浮出水面。

“这不仅是美国的问题。英国在这场斗争中也在扮演狠角色，英国政府通信总局的作为甚至比美国还恶劣。”斯诺登对英国《卫报》说。据报道，英国情报机构政府通信总部 GCHQ 在过去一年半的时间里，对多条承担国际电话和互联网信号的光缆系统进行秘密监控，不但拦截和存储了海量的私人通话、电邮、浏览记录等数据，还与美国国家安全局彼此共享信息。

值得注意的是，GCHQ 可对从光纤获得的海量数据储存 30 天并进行分析。这一行动的代号为“Tempora”，2011 年正式投入使用。监控光纤的技术能力使得英国政府通信总局成为情报界的超级力量。该机构的互联网监控能力是美国、英国、加拿大、澳大利亚、新西兰五大监视窃听联盟成员中最强的。

“Tempora”事件曝光后，英国相关部门的辩解与美国情报部门的说法何其相似：这些数据收集都有监管机制，它们中的许多信息对及时发现和防范重大犯罪颇有功劳。为了国家安全而进行的秘密监控是否应该公之于众？在反恐监控和保护公民隐私上，如何做到更符合民主性要求的平衡？**对于这些问题，像爱德华·斯诺登一样的自由主义者更加相信：透明与个人隐私才是自由社会的基石，保密与监督是暴政的大门。**

在全球舆论关注和疑虑中，斯诺登在互联网上留下了他的回答：“我不希望上帝知道我是谁。”

隐私的边界在哪里

“互联网上的一切免费平台、免费服务和免费内容都需要定向广告来买单”，亚历山大·福尔纳斯在《大西洋月刊》上曾坦言，“而定向广告的效力及其盈利程度又将依赖于对用户数据的搜集和整理。”

Google 使用我们的个人数据搜集并分析处理出不同种类的价值信息，最终都体现到了市值里。这么做的公司当然不只有 Google 一家，可以说这是目前几乎所有互联网公司的普遍做法（具体他们是如何凭借用户数据获得收益的算法在之前的章节曾做介绍）。这是一场几乎没有人明确点头的交易，不论你是否愿意，都已参与其中。**互联网公司免费使用我们的各种数据，并赚取丰厚的收益，同时作为交换，提供给我们那些看似“免费”的服务。**

为什么我们如此在乎自己的隐私，却会将自己认为如此重要的隐私以很低的成本出卖？美国著名的波士顿咨询公司（BCG）曾在 2013 年针对人们如何看待个人数据隐私这一问题在多个国家进行了调查。调查中，个人数据被分为姓名、年龄性别、财务、子女、健康、通话记录、社交网络信息等多个项目，各国民众普遍认为个人财务信息属于最为隐私的内容，其中中国和日本的民众表现出来的隐私观念要比其他国家更淡薄。对于大多数的个人数据项目，日本受访者基本只有 10％～30％认为属于个人隐私，即便是公认的最为敏感的个人财务信息，也只有 34％的日本受访者认为这是隐私。具体结果如图 7.9 所示。

也就是说，人们在判断什么数据属于个人隐私上是存在差异的。多数人并不知道有多少关于自己的数据正在被收集，以及其潜在代价和收益是什么。相反，当你的隐私数据被善意利用时，会让你感觉如同享受到了“私人定

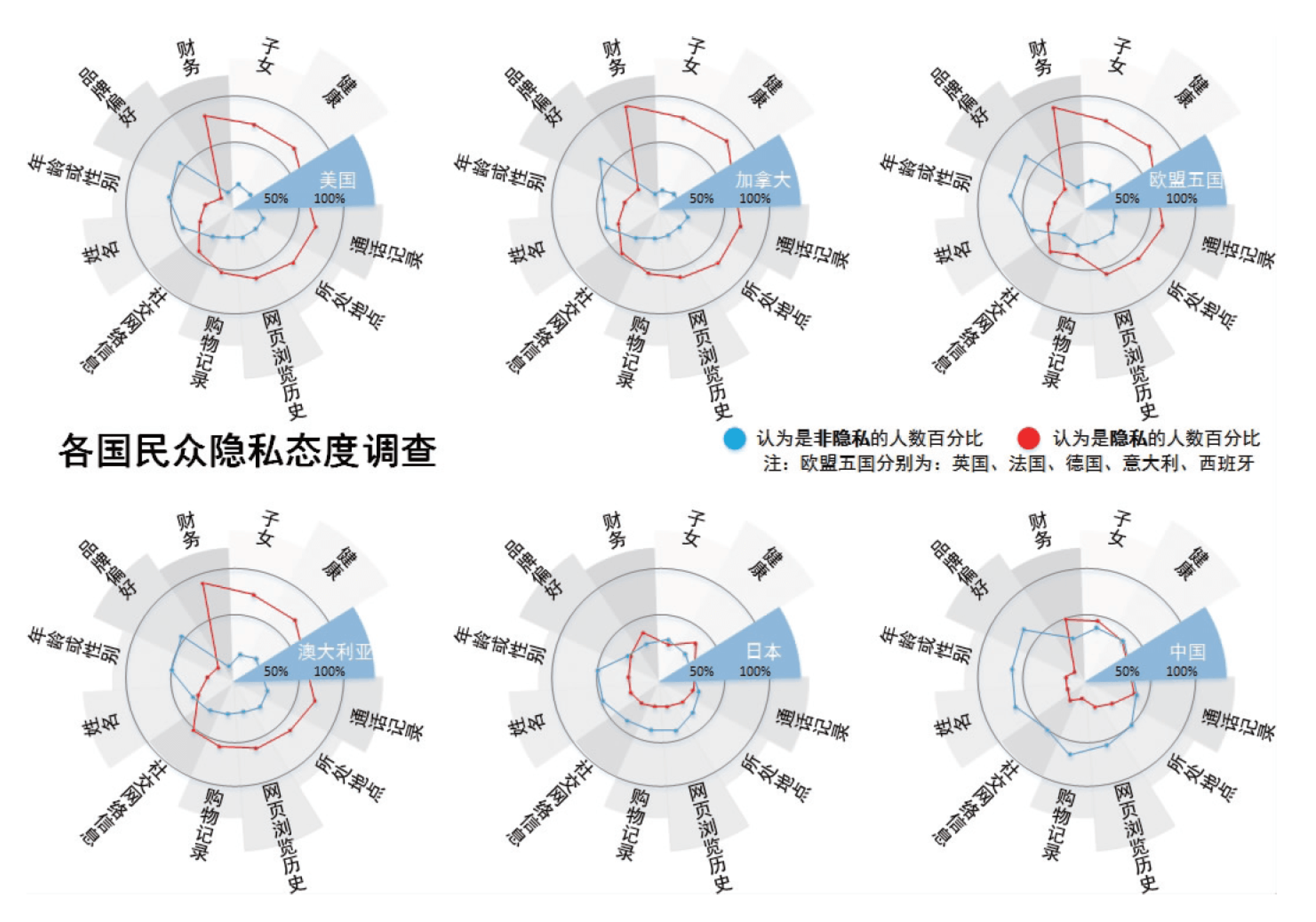


图 7.9　各国民众关于个人隐私问题的调查结果分布情况

制”的服务，完全不知道越来越多地暴露自己的隐私数据其实也是一件存在风险的事情。

在“棱镜门”事件曝光后，人们也开始重新审视网络安全和个人隐私问题。2014年，一项由约瑟夫·朗特里改革基金会发起的调查显示，85%的网民认为，上网浏览记录等信息的保密工作“相当重要”；英国市场调研公司Mori的调查也显示，仅有12%的受访者认为这些隐私是否被监控无所谓。与上述两项数据形成对比的是，美国大规模监控计划刚被揭露的时候，仅有约40%的网民在意自己的隐私是否被监控。

尽管民众针对“网络隐私保护”的呼声甚嚣尘上，但结果往往像是一个钟摆随着不同的时期事件左右晃动。2015年12月，美国加州圣伯纳迪诺出现了恐怖袭击，28岁的赛义德·法鲁克和他29岁的妻子塔什芬·马利克对一家社会服务机构发起袭击，造成14人死亡。随后，两人在同警方的枪战中死亡。联邦执法人员在他们的汽车上找到一部iPhone手机，但是苹果公司拒绝帮助解锁这一嫌犯手机。虽然联邦法官匹姆之后裁决要求苹果公司为联邦调查局提供能够解除手机安全功能、避免多次解锁不成功自动消除数据的方法，但苹果CEO库克表示，美国政府对iPhone手机的要求是危险的，这无异于美国政府在要求围绕苹果手机加密功能创建一个“后门”，并指出这样的要求“没有先例”。

美国独立民调机构皮尤中心针对“苹果对抗FBI”事件做了一项调查，结果发现超过一半的美国民众支持FBI解锁枪击案中的iPhone。而网络调查公司SurveyMonkey的调查结果几乎与其一致，有趣的是，这份调查还显示，只有16%的受访者读过库克针对此事发表的公开信，但在读过该公开信的受访者中，超过50%支持苹果（图7.10和图7.11）。当时许多媒体在报道SurveyMonkey的调查时，几乎不约而同地强调了是否读过公开信这一区别，不免让人思索：对苹果解锁iPhone的后果有更深理解的人，是不是就更

容易站到苹果这一边？换言之，是否不支持苹果的人只是因为对隐私缺乏起码的认知？



图 7.10　加州圣伯纳迪诺枪击案后，民众支持苹果表态

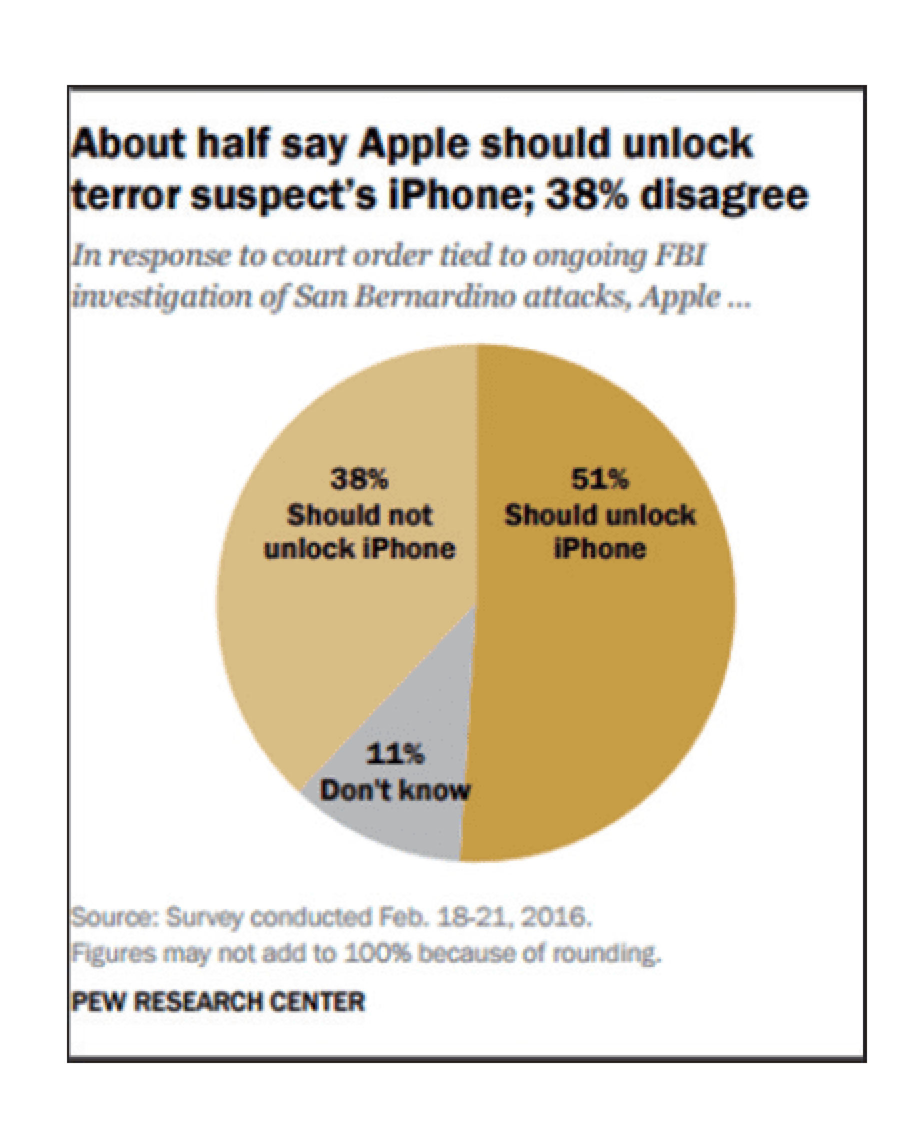


图 7.11　51% 的受访者支持 FBI，仅有 38% 支持苹果

明知隐私可能受到侵犯却依然要加入 FBI 们的阵营，以及明明介意政府窥探自己的隐私却在保护隐私的行动上表现出迟钝，这些似乎都暗示了一个问题：人们并不是不在乎自己的隐私可能泄露，而是，当他们权衡隐私与安全或是其他因素时，不仅无知也时常抱有一种侥幸心理。在苹果与 FBI 的争执中，硅谷大佬挨个站队，17 家科技公司向法庭联名递交法律文件声援苹

果，科技公司不断强调给政府开后门可能出现的“滑坡效应”[①]。如果说，科技公司关注的是，“万一解锁技术落入不法分子之手会怎么样”或是“万一政府提出进一步的监控要求该怎么办”，那么普通民众关心的可能就是，“我会不会成为那个万分之一的受害者。”

2016年皮尤又做了一项调查，结果发现尽管担心隐私的人比不担心隐私的人略多一些，但真正从行动上保护个人隐私的人却很少。在所有听说过政府监听计划的受访者中，仅有1/4的人表示他们因此改变了使用各种技术的方式，比如更改隐私设置、卸载某些应用或定期修改密码，如图7.12所示。

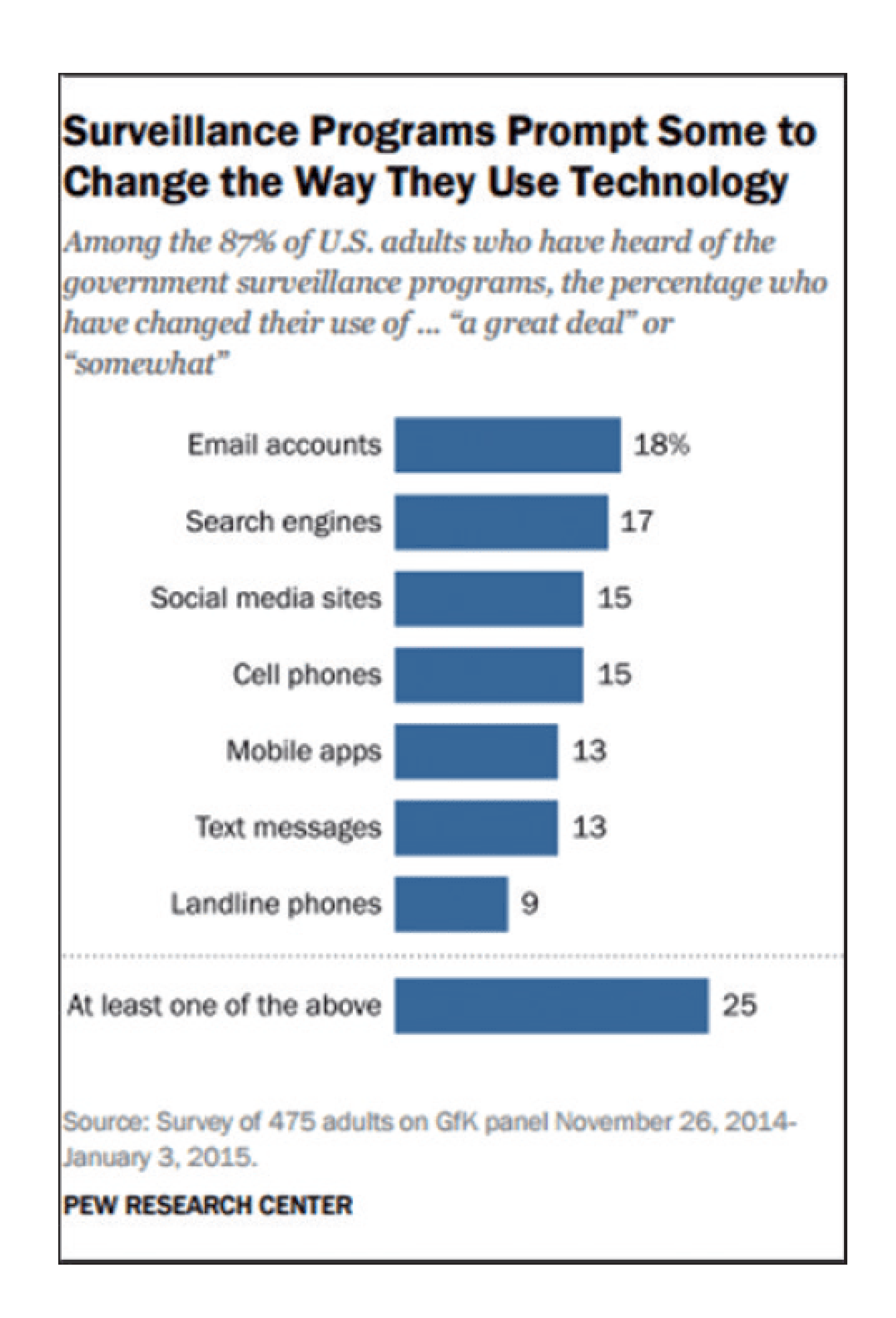


图7.12 约1/4受访者表示在“棱镜门”后改变了使用技术的方式

① “滑坡效应”(Slippery Slope)在逻辑学中是指：如果我们今天允许一件相对无害的事情发生，我们可能会启动了一个趋势，最终结果是，我们眼下认为不可想象的事情也会发生和被接受。

对于隐私边界的模糊认识，以及对隐私保护技术所涉及的知识和成本（比如更换聊天工具、安装反监控软件）的缺乏了解，使得人们在面对隐私安全问题时更加脆弱。

尽管我们是无足轻重的小人物，但我们的隐私真的将变得无处遁逃吗？

在觊觎中寻求算法保护

“大富翁”（Monopoly），又名“地产大亨”或“强手棋”，是一款经典的多人策略型棋盘游戏。游戏中，参与者需要掷骰子前进，同时使用多种道具、卡片，通过买地、建楼、收取过路费等方式来与对手进行财富比拼。坚持到最后，能迫使其他对手破产的玩家获得胜利。

据估计，自 1904 年最初版本推出以来，“大富翁”是世界上最多人玩过的棋盘游戏，人数可能超过 6 亿。但鲜为人知的是，在第二次世界大战中，它还承担了一项秘密使命，英军正是依靠大富翁棋，从德军手中营救出无数的英军战俘。可以说：“没有大富翁，就没有现在的英格兰。”

——詹姆斯 • 于勒（James Yule），英国陆军中校

在第二次世界大战中，德军有 20 多万英军俘虏，其中包括大批飞行员。当时飞行员非常稀缺，所以英国秘密情报局成立了一个战时组织“军情九处”，专门帮助营救英军战俘，特别是飞行员。

遵照日内瓦公约，当时德国允许英国慈善基金会等公益组织向战俘营运送食品、衣物等物资，其中也包括娱乐游戏和棋牌等。英国军情九处曾通过这些组织多次向战俘营里输送逃生工具，比如藏着收音机元件的棒球，藏有地图的黑胶唱片，藏在鞋跟里的小刀，伪造身份证等，结果都被德军一一识

破了。

“大富翁”游戏中设置了各种各样的游戏道具，尤其是包含地图、钱币等，对于战俘逃跑，工具和地图都是必要物品，所以英国情报局注意到了这款游戏，并开始投入大量精力进行重新修改和设计。首先是如何绘制用于逃跑的地图，采用纸质地图不容易保存、遇水容易破损，如果展开纸质地图也会发出较大的响声，容易被德军发现。于是，英国情报局想到了使用丝绸，1941 年英国情报局与擅长丝绸印刷的瓦丁顿公司（Waddingtons）秘密达成了协议，选择了一处狭小、安全的车间，只有极少数员工参与，各自分工将一切秘密信息以及所有逃生出口都印制在了游戏地图当中，第一款特别版“大富翁”游戏棋诞生了。

这款特别版游戏里，不仅有标准的游戏道具，例如骰子、房屋模型、机会卡和苏格兰犬，还增加了暗藏的金属锉刀、指南针等特殊道具，其中的丝绸地图一并标注了逃生路线上所有可供躲藏的安全场所。另外，各套特别版游戏棋所装入的特制道具略有不同，有的游戏棋中通过一些马太福音真言，巧妙地传递着情报信息（图 7.13），甚至还有一些游戏棋里，部分游戏币被替换成了德国、意大利和法国等国家的真实货币，以保证战俘在逃跑途中可以使用。



图 7.13 战俘营中正在进行大富翁棋游戏

借助与多家英国当地慈善基金会合作,英国情报局慢慢地将特别版大富翁游戏(棋盘上停车场的右下角有个红点)混入普通版游戏棋中一起送入了德军战俘营。为了防止这一逃生工具再次被德军发现,监狱里秘密组成一个"逃跑委员会",收到各种逃跑工具后大家都先交给逃跑委员会,委员会只负责保管工具并不会询问工具的来源,谁计划逃跑就可以向逃跑委员会申请工具。这样做的目的是为了避免有人叛变,叛徒也不会知道工具到底是隐藏在了什么物资里。**就这样,特别版大富翁游戏成了二战期间唯一没被德军识破的情报工具**。据二战结束时估计,从德军战俘营中成功逃出的战俘超过 35 000 人(图 7.14)。



图 7.14　大富翁棋与被解救的英国飞行员

这段惊险悲壮的逃生之旅,不仅巧妙地把各种工具隐藏在大富翁棋中,而且通过切断大富翁棋与逃跑工具之间的关联,使得大富翁棋像一般物品一样在诸多物资中隐匿起来,保护了重要的情报信息。这些在军事上经常使用的情报隐藏方法也为今天的数据隐私保护提供了思路。

把隐私匿名

爱德华 · 斯诺登第一次写邮件给《卫报》记者格伦 · 格林沃德(Glenn

Greenwald,曾报道过美国“棱镜”计划)时,使用了一种 PGP 邮件加密软件。不过,据《连线》网站介绍,斯诺登还使用了另一项匿名技术,即一个名为 Tails 的便携口袋式匿名操作系统(图 7.15)。



图 7.15 Tails匿名操作系统

Tails 匿名操作系统[①]不会在本地存储任何用户数据,防止遭受恶意软件攻击,最重要的是 Tails 可以保护数据信息的来源。没有人知道 Tails 的发明者是谁,Tails 的开发成员们一直保持匿名,不仅保护了自己的身份,同时也保护了 Tails 的代码避免受到美国国家安全局(NSA)的施压和干涉。根据 Tails 开发成员们的说法:“那时我们已经是匿名网络 Tor 的爱好者,并且参与开源自由软件社区好几年。Google、Facebook、Yahoo 等网络巨头和国家情报机关为了自己的利益,希望每个人的生活在网络上变得越来越透明,我们想要对抗这种威胁个人隐私的趋势。我们发现网络上缺少的是一个工具箱,可以将所有保护隐私的加密工具整合起来,让普通用户也能够很方便地使用。”斯诺登泄露的一份文件中显示,NSA 当局在一个幻灯片里曾抱

① Tails 匿名操作系统的开源网站:https://tails.boum.org/。

怨 Tails，因为该匿名操作系统对 NSA 不利，这也意味着 Tails 对隐私保护是有用的，任何人如果对 Tails 存有怀疑，可以随时进行检查，所有 Tails 操作系统的代码都是开源的。

最初，Tails 匿名操作系统来自一个名为 Amnesia（健忘症）的软件项目，随后与一个已有操作系统 Incognito（隐身出行）融合形成了 The Amnesic Incognito Live System，首字母缩写为 Tails。目前，Tails 除了有核心小组在对笔记本和台式机的桌面操作系统进行研究，还有独立小组在 Android 和 Ubuntu 平板上开发移动版本操作系统。实际上，Tails 是一个匿名处理的完整 Linux 操作系统（主要是基于 Debian GNU/Linux），它的设计独立于计算机原有操作系统，可以安装在 DVD 光盘、U 盘和 SD 卡上，不仅支持随身携带，而且需要时可以直接用 Tails 启动计算机，之后计算机联网会自动运行 Tor 匿名网络。如果关闭 Tails，计算机将重新启动其原有的操作系统。

Tails 可以帮助我们在线匿名使用互联网，同时它的配置特别注意不去使用计算机的硬盘，即使它们之间有一些交换空间也是如此。Tails 使用的唯一存储空间是内存 RAM（断电自动擦除存储内容），所以 Tails 允许我们在任何计算机上使用敏感数据信息，比如在酒店的计算机或图书馆的计算机上进行操作，我们不会留下任何数据使用的痕迹或任何存储记录，并在关闭计算机后避免我们的数据被恢复。Tails 的这种性质被称为“失忆”，能够更好地保护个人隐私，如果我们明确地要求它，就可以将特定的文档显式存储到另一个 U 盘或外部硬盘上，并将其带走以备将来使用。纪录片导演劳拉·珀特阿斯（Laura Poitras，拍摄的《第四公民》（*Citizenfour*）获得 2015 年第 87 届奥斯卡最佳纪录长片奖，图 7.16）使用 Tails



图 7.16　第 87 届奥斯卡最佳纪录长片《第四公民》

匿名操作系统时说道："虽然 Tails 的安装和验证略有麻烦，一旦完成了设置再使用起来是非常简单的。"

Tails 作为一个隐私保护的工具箱，预先配置了许多加密程序和匿名工具，最为知名的就是匿名网络 Tor。Tor 是由全球志愿者共同组成的匿名计算机网络，它允许用户以完全匿名的方式浏览及访问互联网，帮助用户完全隐藏网络使用记录且不受地域限制。现在使用 Tor 匿名网络的用户很多，大部分是想要保护隐私的普通人，但也包括了非法交易的成员、极端分子、军方、通缉犯等，任何工具在保护普通人的同时也都会被犯罪分子利用，这似乎是不可避免的，不能因为刀可以用来实施犯罪，就不让普通人买刀啊！

洋葱网络——一个谷歌看不见的世界

Tor 的核心算法是"洋葱路由（The Onion Router）"，在 20 世纪 90 年代中期，由美国海军研究实验室的数学家保罗·西维森（Paul Syverson）、计算机科学家迈克·里德（G. Mike Reed）和大卫·戈尔德施拉格（David Goldschlag）发明，用于对美国的数据情报通信进行保护。现在的 Tor 项目是自由软件，并由一个位于马萨诸塞州的非营利组织 The Tor Project 进行维护。

2015 年 10 月，自由程序员卢克·米尔兰达（Luke Millanta，来自悉尼）通过一个名为"Onionview"的计划，分析了 Tor 匿名网络的现有规模和发展情况并感叹："人们谈及 Tor 网络以为是一小撮使用者在地下室里隐蔽操作着计算机实现的，如果他们看到 Onionview 地图结果就会发现，'天啊，全球看上去至少已经有 6000 个 Tor 网络节点了。'"其中，德国的总节点数有 1364 个，已超过了位居第二的美国（1328 个），如图 7.17 所示。

2016 年初，数据可视化软件公司 Uncharted 也发布了"TorFlow"计划的观测结果，除了德国和美国，拥有较多 Tor 节点数量的国家还有法国、荷兰、

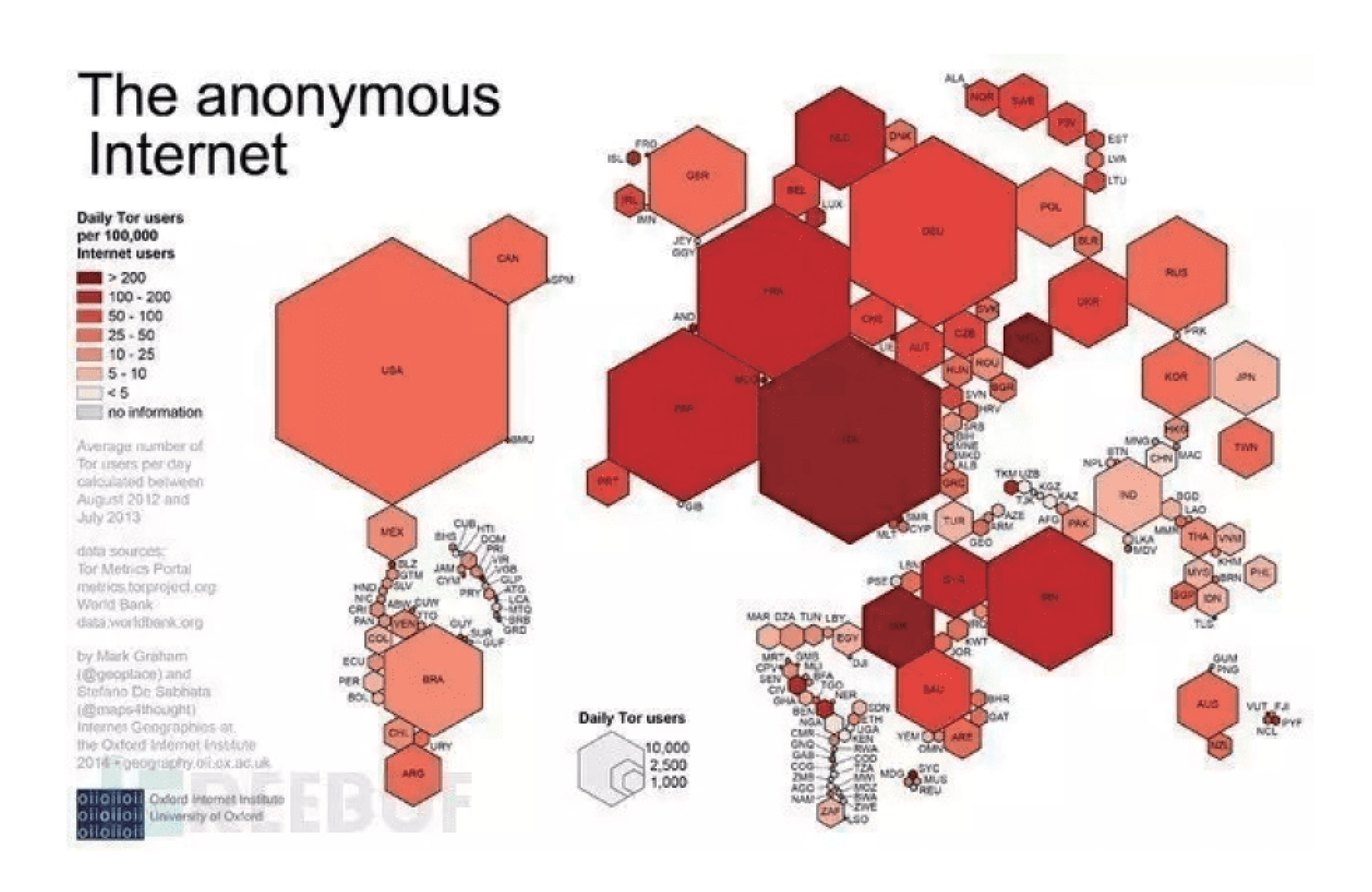


图 7.17　Tor 匿名网络每日用户数量实时地图

英国和俄罗斯(图 7.18)。TorFlow 还显示出了一些我们料想不到的 Tor 节点位置,例如利比亚和利比里亚,这里不仅有民众在使用 Tor 网络,也存在着 Tor 志愿主机节点。6 年前,Tor 匿名网络还只有 2000 个主要节点,如今 Tor 网络已经彻底去中心化并由分布于全球的 6425 个节点组成。每年下载 Tor 软件的用户近 5000 万人次,连 Tor 的发明者们都承认,自己也无力摧毁 Tor 了。





图 7.18　TorFlow 计划于 2008 年(左)和 2016 年(右)观测到的 Tor 节点数量和变化趋势

Tor 在短时间内成为全球应用最为广泛的匿名网络之一，不仅仅在于它能够隐藏用户访问网络的真实身份，它同样使得各类网络服务提供者可以隐藏自己的身份和服务器所在的位置。某个 Tor 使用者可以通过 Tor 隐匿服务功能建立网站，其他 Tor 用户也可以在网站上发布资料信息，他们不用担心会被追踪或遭到审查，因为没有人能够确定到底是谁在运行网站，即使是网站运行者也不知道到底是谁发布了那些数据。例如，中东地区的 Tor 用户创建了一个网站转载西方媒体的新闻报道，可以使用 Tor 软件、Tor 的主机，借助 Tor 网络“介绍点”(introduction points)来引导所有想要访问网站内容的人，以介绍点的地址连接这个隐藏的网站，但是没有人能找到其真实位置，网站访问者互相也都不知道对方的真正网络标识。这就是所谓的“洋葱网站”，常规的搜索引擎是无法到达这样的网络空间的，它也由此得到一个更为耸人听闻的名字“暗网”，Tor 为“暗网”提供了一个秘密掩体，因此 Tor 也被认为是开启了一个“谷歌看不到的世界”。

Tor 匿名网络通过一系列中间计算机又叫中继节点(即志愿者计算机)，在我们的终端计算机和访问目的地之间建立起一系列的跳跃式连接，简单来说，Tor 是想要用一条拐弯抹角的、难以辨认的路径甩掉任何跟踪我们的人(图 7.19)。当用户启动一个 Tor 代理客户端时，它将会从 Tor 的目录服务器获取 Tor 网络的节点列表。接下来，客户端会在网络中任意选择一跳中继节点建立加密连接，该加密连接会在匿名网络中逐跳扩展，直至到达目的地。例如，用户 Alice 通过三个中继节点访问目的地 Bob，如图 7.20 所示，实线连接箭头表示加密通信，而虚线连接箭头表示一般未加密连接。在 Alice 和 Bob 的通信中，每个中继节点只知道自己从上一跳节点接收数据以及向下一跳节点发送数据，其他部分都是加密不可见的。这就意味着第一个节点只知道连接的发起者而不知道目的地，最后一个节点只知道连接的目的地而不知道发起者，没有单独一个中继节点会知道数据的完整传输路径。这样一来，即使攻击者使

用嗅探技术或直接攻击了某个中继节点，也不会获得这次连接的完整信息，难以实施消息和来源的追踪。最后，如果用户 Alice 再一次访问 Bob 或者另一个目的地 Jane，客户端将会重新选择中继节点建立不同的连接路径。

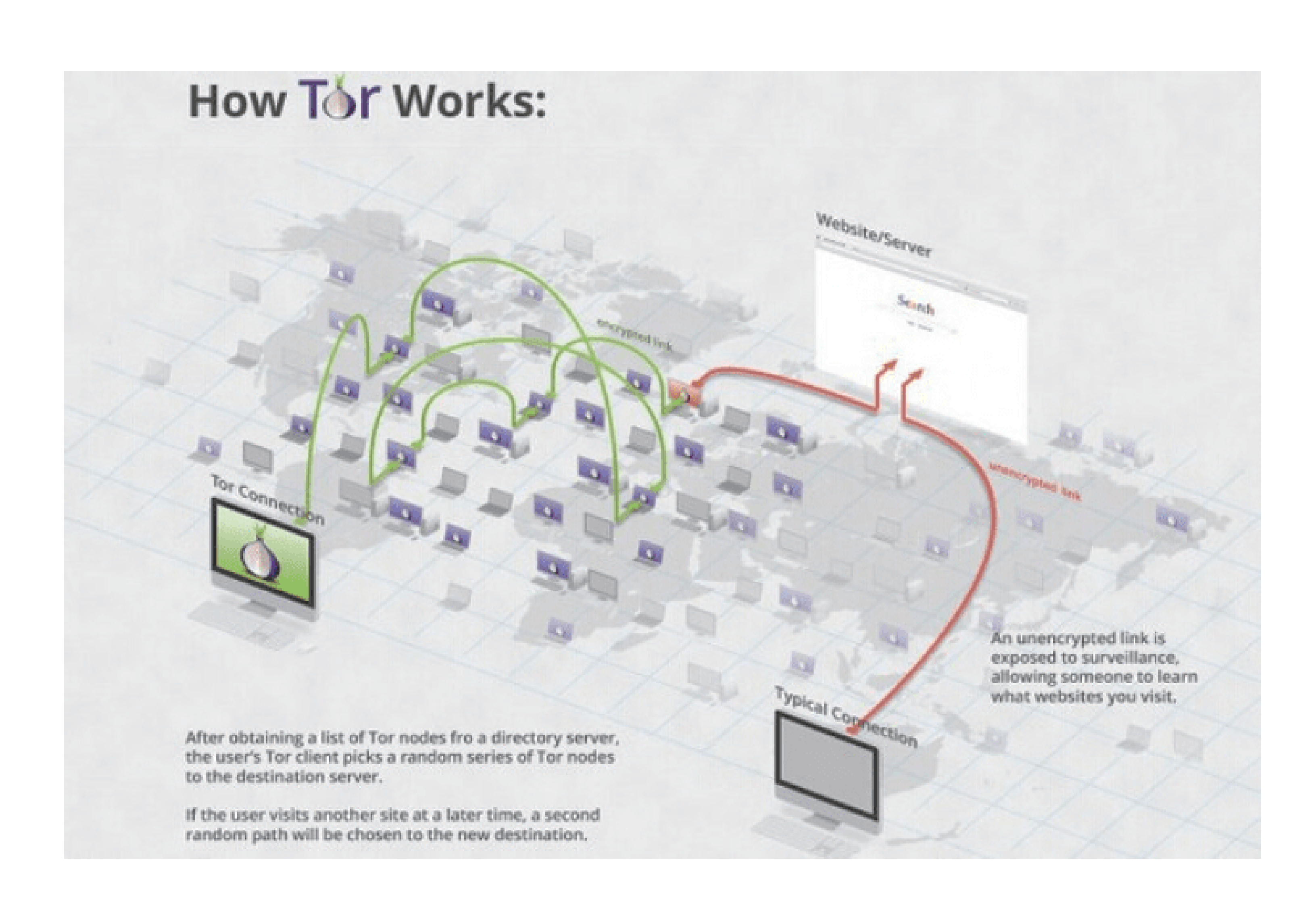


图 7.19　Tor 匿名网络创建连接示意图

洋葱网络，顾名思义，用户的客户端在发送数据时，会对 Tor 数据包进行层层加密封装形成像洋葱一样的结构。一个数据包经过 Tor 网络的各个中继节点传输，其头部被一层一层解密剥去，最后一跳出口中继节点会完成最后一层解密，得到真正的目的地信息。

由于 Tor 路由算法采用了多重路由的模式，其网络传输效率会受到一定的影响，为了减少性能开销，Tor 网络在建立连接时使用 RSA 加密，而在传输实际数据时使用速度较快的 AES 加密。构成中继节点的志愿者计算机，并不需要运行特别的硬件，只是配置 Tor 软件，他们的计算机会贡献一部分带宽资源支持 Tor 网络运行，如果加入 Tor 网络的志愿计算机越多，那么 Tor 网络的传输能力会更强，防止 Tor 的用户信息被追踪的效果也会更好。针对一些关于 Tor 网络是否会产生危害的争论，Tor 项目主页还发表了一项“为什么 Tor 不

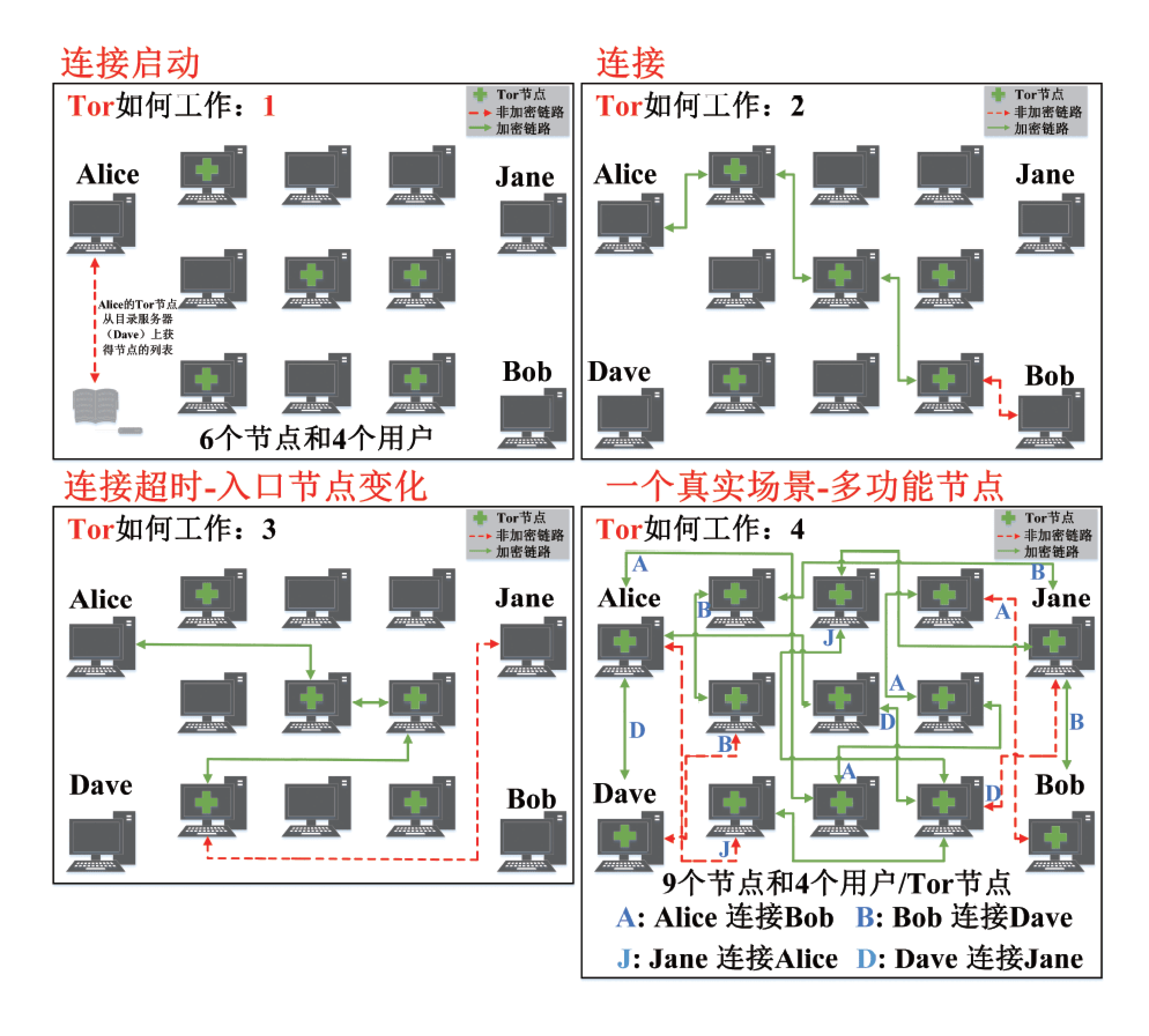


图 7.20　洋葱路由算法工作原理示意图

会帮助犯罪”的精彩声明，Tor 的项目组成员们相信，“Tor 的存在对犯罪行为的帮助并不会比 Internet 多。”这句话笔者个人是非常同意的。

除了 Tor，Tails 匿名操作系统上还装载了其他的加密工具，如 PGP、密码管理系统 KeePassX、聊天加密插件 Off-the-Record 等。这些应用都经过修改，并提升了安全性。当然，没有操作系统和加密工具可以保证完全的私密。

尽管 Tails 中有一些办公应用，如 Open Office、GIMP、Audacity，但它并不适合在我们的日常网络活动中使用，由于平日里我们总是需要使用一些与自己身份有关的网络服务，这样会将我们完全暴露。因此开发者们建议用

户：应该在一些需要保持匿名的特殊活动中使用 Tails。

被算法设计的隐私保护

夏洛克：我昨天第一次见到你时问你是去过阿富汗还是伊拉克，你好像很惊讶。

华生：对。你到底怎么知道的？

夏洛克：你的发型、站立姿势说明你是军人出身。但你说的话有点时过境迁，说明在巴茨医院受训过，所以显然是军医。脸上晒黑了但手腕却没有，说明你从国外回来，但不是日光浴。你走路跛得厉害，却宁愿站着也不要求坐下，所以多少有点心理疾病，导致你创伤的最初原因，应该是在战场上受伤。战场加上晒黑，那就是阿富汗或伊拉克。

——《神探夏洛克》，第一季第一集

夏洛克·福尔摩斯（Sherlock Holmes）与约翰·华生（John Watson）在第一次相遇时，曾使出了浑身解数只为了让对方喜欢自己，对于即将合租同一幢公寓的华生，夏洛克习惯性地给出了自己恰到好处的推理，这些推理都与华生并不想向外公开的实际情况相吻合。**现实生活中，如果没有高智商和大量的背景知识，短时间内做到夏洛克这样的完美推理似乎很难，但是有了互联网、有了大数据相关技术，找到那些隐藏起来的事情，似乎又变得简单起来。**

下面给个例子，网友仅仅根据王珞丹在微博中贴出的图片和说过的话就推断出她曾经居住的地址。

第一步：信息获取；素材：两张照片。

网友从王珞丹的博客和微博中选取了两张图片，一张是从窗内俯瞰小区绿化植物的照片，一张是从窗内拍摄的窗外全景图。网友认为照片透露了几个主要信息：第一，楼体外观和窗框难擦干净的痕迹，说明这是已经建成一段时间的西式小区。第二，王珞丹家在高层。第三，小区内有三个在一条直线上大小一样的正方形花坛。

第二步：区域筛选；素材：4 条微博留言、电子地图。

王珞丹微博 1："四环堵死了，联排迟到了。"这说明，她家在四环外。

王珞丹微博 2："演出这么多年，还没有在北京中心地带买一套房子。"这说明，说明她家不在市中心。

王珞丹微博 3："爸爸送我和小 6 去给《无人驾驶》配音，光顾着看微博留言，忘记给老爸指路，车都开到中关村了！"（爸爸开始唠叨我说开导航吧）这说明，她家不在中关村及进城路过中关村的地方。

王珞丹微博 4："患了严重的痢疾，20 分钟后赶到了附近的一所小医院。"这说明，她家周边无大医院。

网友继续将北京城区地图划为 9 个区域。根据微博留言用排除法分析小区所在地。

第三步：网络搜索；素材：电子地图。

网友在电子地图上截取锁定区域俯视图，放大局部寻找王珞丹照片中有标志性花坛的小区，很快就找到目标。

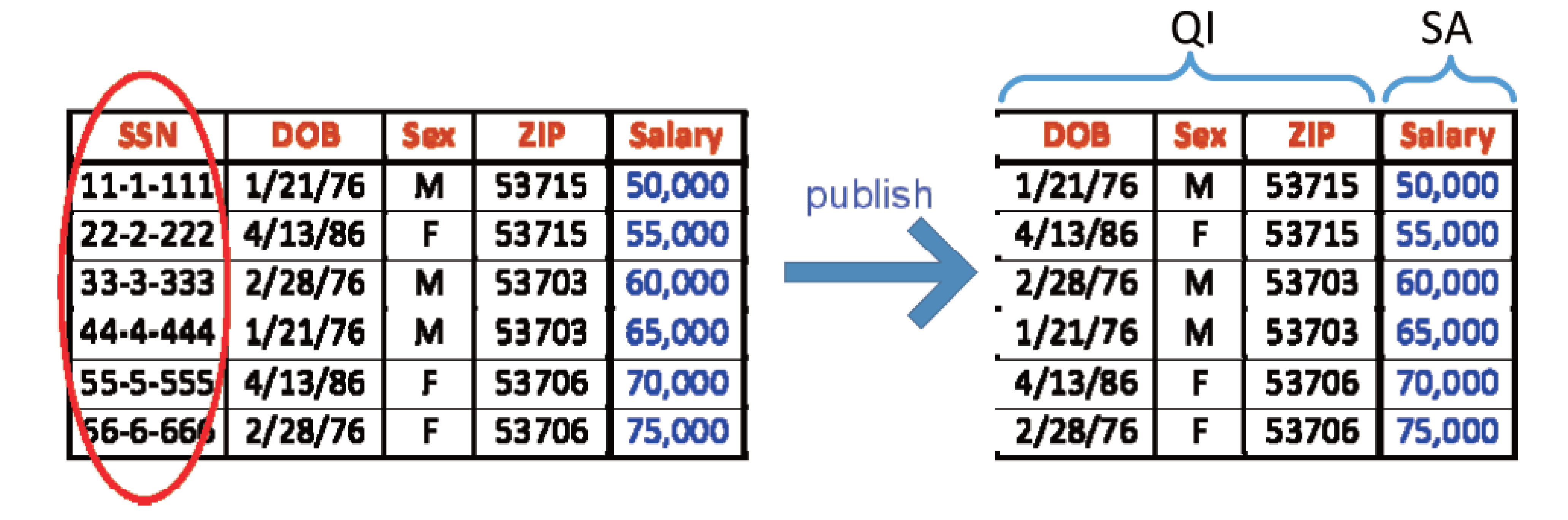
第四步：实地核对；素材：照片。

该网友亲身前往这个小区，现场拍摄照片，与王珞丹所发的照片进行比较，确认推理正确。

如果开始推理

由于网络服务本身的技术特点以及我们日常的很多网络活动自然带有身份标识，关于我们的这些网络行动就自然沉淀为数据。对于这样的数据，我们是否还可以借助匿名化的方式在不同的数据应用接触到数据之前实现隐私保护？另一方面，对于提供各类“免费”服务的网络公司来说，真正有价值的是大数据中某个群体的统计特性而非某个特定用户的具体信息，很多著名的隐私保护匿名算法也是在这一背景下产生的。那么问题来了，如果能在防止隐私数据泄露的同时又能保持数据集的可用性就完美了，是否存在这样的算法？

匿名化的算法思想其实由来已久，最开始一个简单的想法就是去掉数据集中那些起到唯一标识作用的信息，例如，美国某个人口普查记录当中（图 7.21 左）包含社会保险号（SSN）、出生日期（DOB）、性别、邮编和工资收入的信息，如果我们将起到唯一标识作用的社会保险号 SSN 去掉，对公布出来的数据集（图 7.21 右）进行定义，即包含两类信息：类标识信息（QI）和敏感属性信息（SA）。直观上，凭借上述两类信息，将无法明确获知某个人对应的工资收入值，在一定程度上具有匿名化的作用。但是，在遇到链接攻击



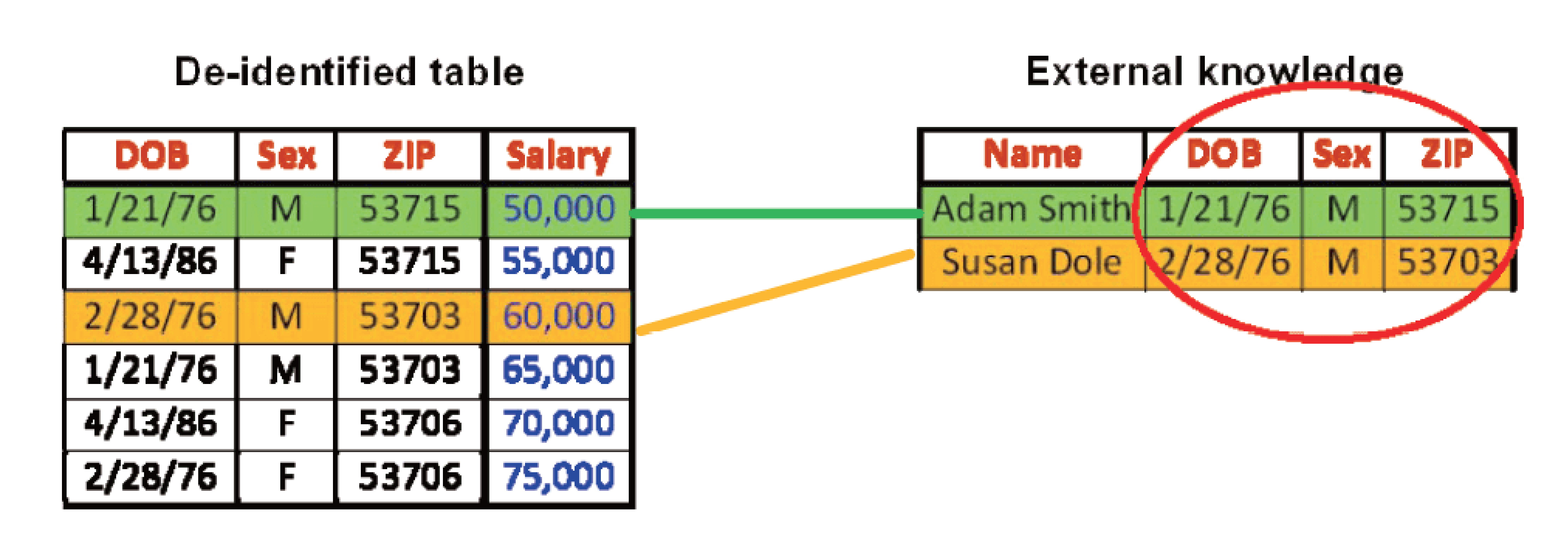
SSN	DOB	Sex	ZIP	Salary
11-1-111	1/21/76	M	53715	50,000
22-2-222	4/13/86	F	53715	55,000
33-3-333	2/28/76	M	53703	60,000
44-4-444	1/21/76	M	53703	65,000
55-5-555	4/13/86	F	53706	70,000
66-6-666	2/28/76	F	53706	75,000

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

图 7.21　美国某人口普查记录表去唯一标识处理

(Linking Attack)情况下，结合不同的背景知识进行推理，这种匿名化作用就十分微弱了。

研究者们发现，以比较容易得到的注册选民信息登记表为背景知识，通过将上述人口普查记录和注册选民信息登记表做链接推理，攻击者就很容易得出大部分在选民信息登记表出现过的人的工资收入值（见图 7.22）。一份研究报告指出，有 87%以上的美国人可以由他们生日、性别和邮政编码的组合唯一确定。一旦攻击者已知 Adam Smith 的生日、性别和邮政编码分别是 1/21/76、Male 和 53715，那么他只要稍加推理就能知道 Adam Smith 的工资收入是 50 000 美元，这样 Adam Smith 的隐私信息仍然被泄露了。



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

Name	DOB	Sex	ZIP
Adam Smith	1/21/76	M	53715
Susan Dole	2/28/76	M	53703

图 7.22　人口普查记录和注册选民信息登记表的链接推理

由此可见，数据发布者既不清楚数据接收者所拥有的背景知识，也难以对数据接收者如何使用数据进行控制，真正实现数据匿名化并不是一项简单的工作。

美国在线 AOL 就曾因匿名化不足而导致数据泄露，这是较早出现的全球重大数据泄露事件之一。2006 年 8 月，AOL 承认将 1900 万条用户搜索查询记录在网上进行过发布，这些信息涉及 65.8 万 AOL 用户以及 1080 多万条不一样的搜索词，时间跨度为 3 个月。AOL 声称，该数据的发布仅作为非商业用途，主要是为了向研究机构提供新的研究工具。对于研究人员来说，

如果要进行搜索记录中的关键词分析，这些真实的来自主流搜索引擎的最新资料可是相当宝贵，AOL 的两名员工就在这些数据的基础上撰写了一份名为“搜索概览”(A Picture of Search)的研究论文。

虽然在公布这些搜索查询记录之前，AQL 已经将用户标识去除，并将同一用户的搜索条目与任意标识链接，但是仍然有一些攻击者成功锁定了某个用户个体，特别是《纽约时报》报出，通过搜索 AOL 第 4417749 号用户搜索记录，最终找到了乔治亚州 Lilburn 市 62 岁的寡妇 Thelma Arnold。Thelma Arnold 进行的搜索查询是：

“现在是访问意大利的最佳季节。”

——AOL 第 441 7749 号用户 Thelma Arnold 的搜索查询

AOL 的这次意外事件发生后，AOL 首席技术官(CTO)莫琳 · 戈文(Maureen Govern)引咎辞职。有关此次泄露事件的一位技术开发人员和他的主管也相继辞职。就在莫琳失去其职位前，她在一次主题演讲中说道，随着互联网的演进，“大公司仍将是大公司。”随后，AOL 迎来的却是倒霉的 10 年。

k-匿名和 l-多样性

为了降低实施链接攻击的可能性，研究人员考虑是否可以通过去除一些类标识 QI 信息或对其进行泛化处理来实现更好的匿名效果。例如，在之前提到的美国人口普查记录表中，我们执行元组抑制(tuple suppression)或 QI 属性泛化(QI attribute generalization)处理，简单地说，就是去除一行的信息或泛化一列甚至多列的信息，原始数据表就变成了缺少部分数据的加工数据表(见图 7.23)。但这样处理的结果是，加工数据表丢失了一些信息条目而损失了数据可用性，却没有很好地保护其他数据隐私(见图 7.23(a))，或者

仍存在部分较为独特的信息条目会被链接攻击查找出来(见图 7.23(b))。由此看来,我们需要考虑的一个关键问题是:数据匿名化到什么程度才能够满足隐私保护的需求?

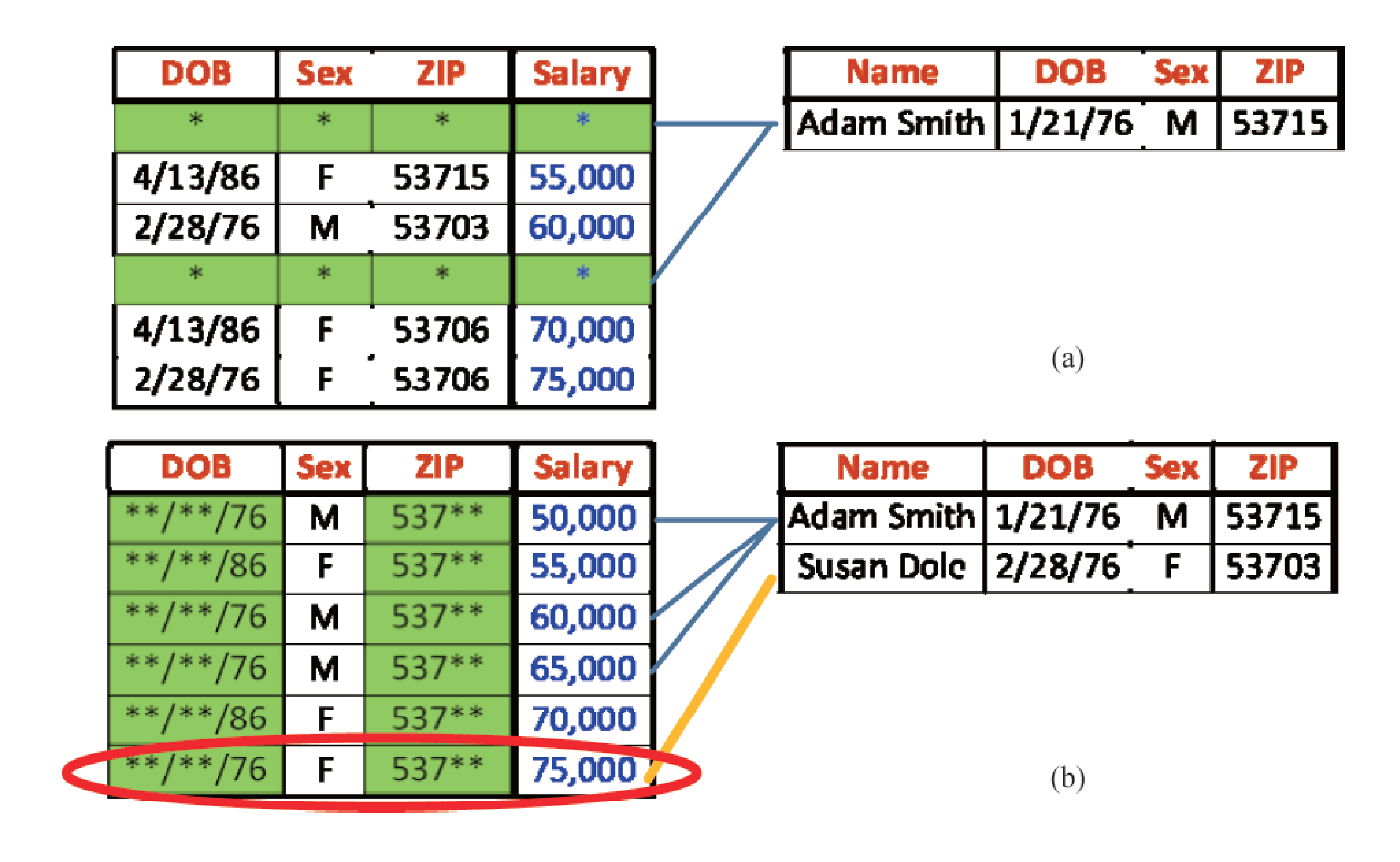
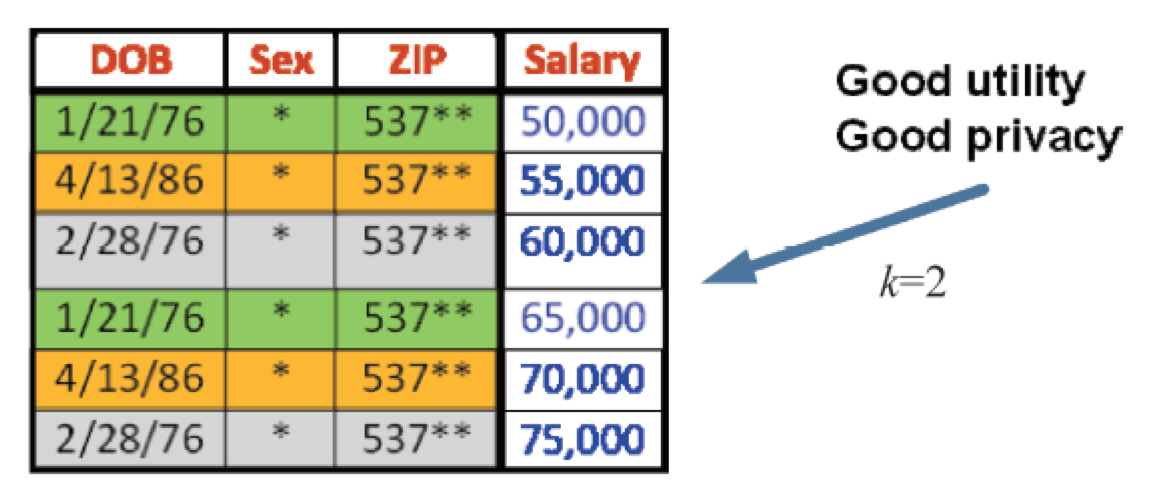


图 7.23　元组抑制或 QI 属性泛化处理后的数据表

通过观察,研究者们进一步发现,对于每条记录在数据表中很可能会与不止一项类标识匹配,是不是只要我们在属性相近的记录之间进行类标识泛化处理,就可以很好地做到隐私保护呢?1998 年,Samarati 和 Sweeney 提出了一种 **k-匿名(k-anonymization)隐私保护算法**,它要求发布的数据中存在一定数量(至少为 k)的在类标识符上不可区分的记录,使攻击者不能判别出隐私信息所属的具体某个个体,从而保护了个人隐私,k-匿名通过参数 k 指定用户可承受的最大信息泄露风险。举例来说,我们对之前提到的美国人口普查记录表再次进行 k-匿名处理,当 k 取 2 时,加工数据表中(见图 7.24)就会存在两条相同的{1/21/76,*,537*}记录,那么{Adam Smith,1/21/76,M,53715}这条记录被精确查找到的可能性就是 1/2。即 k-匿名保证了单独个

体被准确标识的概率最多为 1/*k*。



DOB	Sex	ZIP	Salary
1/21/76	*	537**	50,000
4/13/86	*	537**	55,000
2/28/76	*	537**	60,000
1/21/76	*	537**	65,000
4/13/86	*	537**	70,000
2/28/76	*	537**	75,000

图 7.24　*k*-匿名数据表 (*k*=2)

然而,*k*-匿名真的能够保证信息不会被泄露吗?对于敏感属性,其实 *k*-匿名算法未进行任何限制,如果 *k* 个信息条目被看作一个等价类的话,这个等价类中的敏感属性太过接近也会暴露我们极力想要保护的隐私数据。此时,攻击者仍然可以实施同质性攻击和背景知识攻击,下面我们就给大家展示一下这两种攻击的具体过程。

现在某个医疗机构对外将要发布一份医疗数据,我们划分类标识为邮编(Zip Code)、年龄(Age)和国籍(Nationality),敏感属性为所患病症(Condition)。原始数据表经过 *k*-匿名化处理后(*k*=4)如图 7.25 所示。

Data table of *k*-anonymity, where *k*=4

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	1485*	>=40	*	Cancer
6	1485*	>=40	*	Heart Disease
7	1485*	>=40	*	Viral Infection
8	1485*	>=40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

图 7.25　*k*-匿名化 (*k*=4)处理后的数据表

同质性攻击

假设某一天 Jack 得知邻居 David 突然病倒被送进了医院，于是 Jack 想要猜测 David 到底得了什么病。偶然间 Jack 在网络上获取到了 David 所在医院发布的经过 4-匿名化处理的病人数据信息表。因为 Jack 事先知道邻居 David 是 32 岁的美国人，他们居住的地区邮编是 13064，所以 Jack 很容易发现 David 的医疗记录号应该是 9、10、11 或 12 中的一个。特别是这 4 条医疗记录均显示病人患有癌症，因此 Jack 可以分析出邻居 David 所得的疾病是癌症。

从以上分析过程我们可以看到，采用链接或推理的方法，攻击者可以从数据集中判断出某个目标所在的等价类，假如这个等价类的各记录条目对应的敏感属性值都是一样的，毫无疑问，攻击者就可以清楚地知道该目标的敏感属性值是什么，即目标患者的隐私信息遭到泄露。所以，k-匿名无法阻止同质性攻击。

k-匿名产生的等价类所包含的敏感属性如果缺乏多样性，将直接影响数据隐私安全，但是这种情况并不在少数。我们来做一个粗略估算，假设某一数据集具有 80 000 条不同的记录，并且对应只有 5 种不同的敏感属性（敏感属性与类标识 QI 信息没有关联）。若对这个数据集执行 4-匿名化处理大约可以得到 20 000 个等价类。进一步地，平均每 125 个等价类就会出现一个等价类中所有记录条目的敏感属性值相同，即不具备多样性。因此，上述 20 000 个等价类中将大约有 160 个等价类不具备多样性，相当于 640 个记录对象可能会受到同质性攻击的威胁。由此我们可以看到，即使经过 k-匿名化处理的数据集，拥有相同类标识 QI 信息的记录条目之间，仍然需要保证敏感属性值具有一定的多样性特点。

背景知识攻击

Jack 还有一个微信好友 Aaron，Aaron 在朋友圈中也曾晒出去过 David 就诊的同一家医院，假设 Aaron 的医疗数据信息也包含在已发布的病人数据信息表中（如图 7.25 所示）。Jack 事先已知 Aaron 是 25 岁和日本国籍，Aaron 现在的居住地邮政编码是 13055，由此 Jack 可以初步判断出 Aaron 的医疗记录信息存在于 1、2、3 或 4 中。如果没有更多的参考信息，Jack 此时并不能确定 Aaron 曾患心脏病还是病毒性感染。然而，结合背景知识，日本人在世界人口中患心脏病的比率非常低，所以 Jack 几乎可以推测出 Aaron 的就诊病史是病毒性感染。基于以上分析，k-匿名也不能保证数据信息不会受到背景知识攻击。

为了解决 k-匿名化的数据表仍然会泄露敏感属性值这个问题，2006 年 Machanavajjhala 等人提出了 l-**多样性匿名保护算法**，它要求每个等价类中至少存在 l 个“较好表现的”敏感属性值（即 l 种敏感属性值互不相同）。对 l-多样性匿名算法更一般化的表述是，假设有一个等价类 G，G 中所有元组的敏感属性取值中出现最频繁的取值为 v，出现的次数为 $c(v)$，如果 $c(v)/|G|<=1/l$（$|G|$是 G 中的元组数），那么 G 就满足 l-多样性。按照这种定义，划分后的每个等价类对应的敏感属性值分布将会比较平均，避免出现极端的情况。

这次我们对该医疗机构所要发布的医疗数据表进行 3-多样性处理（如图 7.26 所示），随后把它和图 7.27 作对比，能够看到上面针对 4-匿名化数据表的攻击在这个 3-多样性数据表中并不能奏效。例如，Jack 无法从这个 3-多样性数据表中判断出 David（32 岁、美国国籍、住址邮编 13064）是癌症患者。特别是 Aaron（25 岁、日本国籍、住址邮编 13055）患有心脏病的可能性非常小，Jack 也不能推测出 Aaron 到底是病毒性感染患者还是癌症患者。

Data table of *l*-diversity, where *l*=3

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	<=40	*	Heart Disease
4	1305*	<=40	*	Viral Infection
9	1305*	<=40	*	Cancer
10	1305*	<=40	*	Cancer
5	1485*	>40	*	Cancer
6	1485*	>40	*	Heart Disease
7	1485*	>40	*	Viral Infection
8	1485*	>40	*	Viral Infection
2	1306*	<=40	*	Heart Disease
3	1306*	<=40	*	Viral Infection
11	1306*	<=40	*	Cancer
12	1306*	<=40	*	Cancer

图 7.26　*l*-多样性（*l*=3）处理后的数据表

差分隐私保护

k-匿名和 *l*-多样性算法主要是基于对数据集进行扰动实现隐私匿名保护，这两种算法的不足之处在于没有严格的定义攻击模式，也没办法量化攻击者可能具有的背景知识，导致这两种算法只适用于一些特定场景下的背景知识攻击，在实际应用上存在很大的局限性。

2010 年，Dwork 等人首次提出了差分隐私保护算法，它被公认为是比较严格和强健的保护模型。这一算法与 *k*-匿名和 *l*-多样性算法不同之处在于，它需要基于隐私函数来实现对数据集的保护，它的基本思想是对原始数据、对原始数据的转换或者是对统计结果添加噪音来达到隐私保护效果。该算法可以确保在某一数据集中插入或者删除一条记录的操作不会影响任何计算的输出结果。另外，该算法不关心攻击者所具有的背景知识，即使攻击者已经掌握了除某一条记录之外的所有记录的信息，该条记录的隐私也不会被泄露。差分隐私保护算法最大的优点是，虽然基于数据失真技术，但所加入

的噪声量与数据集大小无关，因此对于大型数据集，仅通过添加极少量的噪声就能达到高级别的隐私保护。

具体来说，给定两个相近的数据集 D 和 D'（后面称为兄弟数据集），二者互相之间至多相差一条记录信息，即 $|D\Delta D'|\leqslant 1$，给定一个隐私算法 A，Range(A)为 A 的取值范围，若算法 A 作用在兄弟数据集 D 和 D' 上任意输出结果 O($O\in$ Range(A))满足下列不等式，则我们认为算法 A 满足 ε-差分隐私。

$$\Pr[A(D)=O]/\Pr[A(D')=O]\leqslant e^{\varepsilon}\approx 1\pm\varepsilon \tag{1}$$

其中，概率 Pr[·]由算法 A 的随机性控制，也代表了隐私被披露的风险。在差分隐私保护算法的实现中，隐私算法 A 构造数据集 D 到 O 的映射时一般需要引入噪音机制，拉普拉斯是一种最为常见的也是最为基本的差分隐私噪声机制之一，通过拉普拉斯分布产生噪音实现对真实值的扰动并最后得到查询返回值。

举个简单的例子，对于某个医疗数据集，当攻击者查询包含 Mary 的 n 条记录中患癌症的人数时（见图 7.27），第一次的返回结果是 35 人，攻击者再执行一次不包含 Mary 的同样记录查询，第二次的返回结果若是 34，那么就泄露了 Mary 患有癌症这一隐私信息，同理返回结果若是 35，攻击者就知道 Mary 没有患癌症。

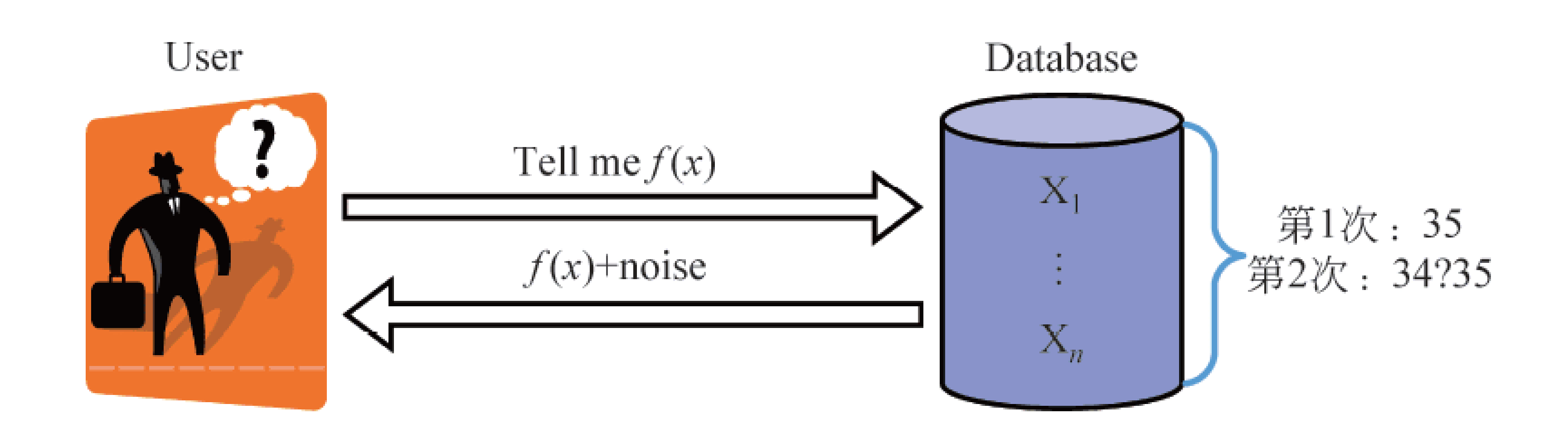


图 7.27　攻击者查询包含 Mary 的 n 条记录中患癌症的人数

经过差分隐私保护，对医疗数据集的查询会出现什么不一样的效果呢？

查询结果会返回一个围绕真实值波动的某种分布结果，分布结果根据加入的噪声的分布情况而定，但是由于有查询成本限制(即次数限制)，不会让攻击者在限制次数内得到整个分布结果，所以攻击者不会获知真实值是多少。

同时，实际返回的查询结果将会满足 ε-不可区分属性，在表达式(1)中，当 ε 足够小趋近 0 值时，兄弟数据集 D 和 D'映射得到的输出结果均为 O 的概率比值几乎为 1，即得到结果 O 的概率几乎相等。也就是表明，不论攻击者的查询请求里是否包含 Mary，两次返回的结果是 35 的概率几乎相等，如果攻击者据此认为 Mary 没有得癌症，就会得到错误的判断，因为 Mary 有可能患有癌症。因此攻击者就再也不能从查询结果中轻易地推断出某项隐私信息了。

对于加入的拉普拉斯噪声，实际上是满足拉普拉斯分布的一个随机值。拉普拉斯分布的概率密度函数表示如下：

$$f(x \mid \mu,b)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{2}$$

其中，噪声是 x-Laplace(μ, b)，位置参数是 μ，尺度参数是 $b>0$。该分布的图像(如图 7.28 所示)是一个尖沙堆形状。

为了保证引入拉普拉斯噪声后，兄弟数据集 D 和 D'关于隐私函数 A 的映射作用仍然满足差分隐私保护的要求，我们取 $b=\Delta f/\varepsilon$，噪声 x 为 Laplace($\Delta f/\varepsilon$)，噪声大小与 Δf 成正比，与 ε 成反比。

之前我们已经知道，ε 越小，兄弟数据集 D 和 D'越难区分，即隐私保护效果越好。这里我们也可以看到，ε 越小，引入的噪声越大，在概率分布函数中反映真实值的最高点相应的会出现尺度和位置上的变化和偏移，当查询预算有限(即查询次数满足隐私保护预算条件下)时，攻击者是无法推断出最高点真实值的。但是数据集的统计特性基本保持不变，带来了较好的数据集可用性。

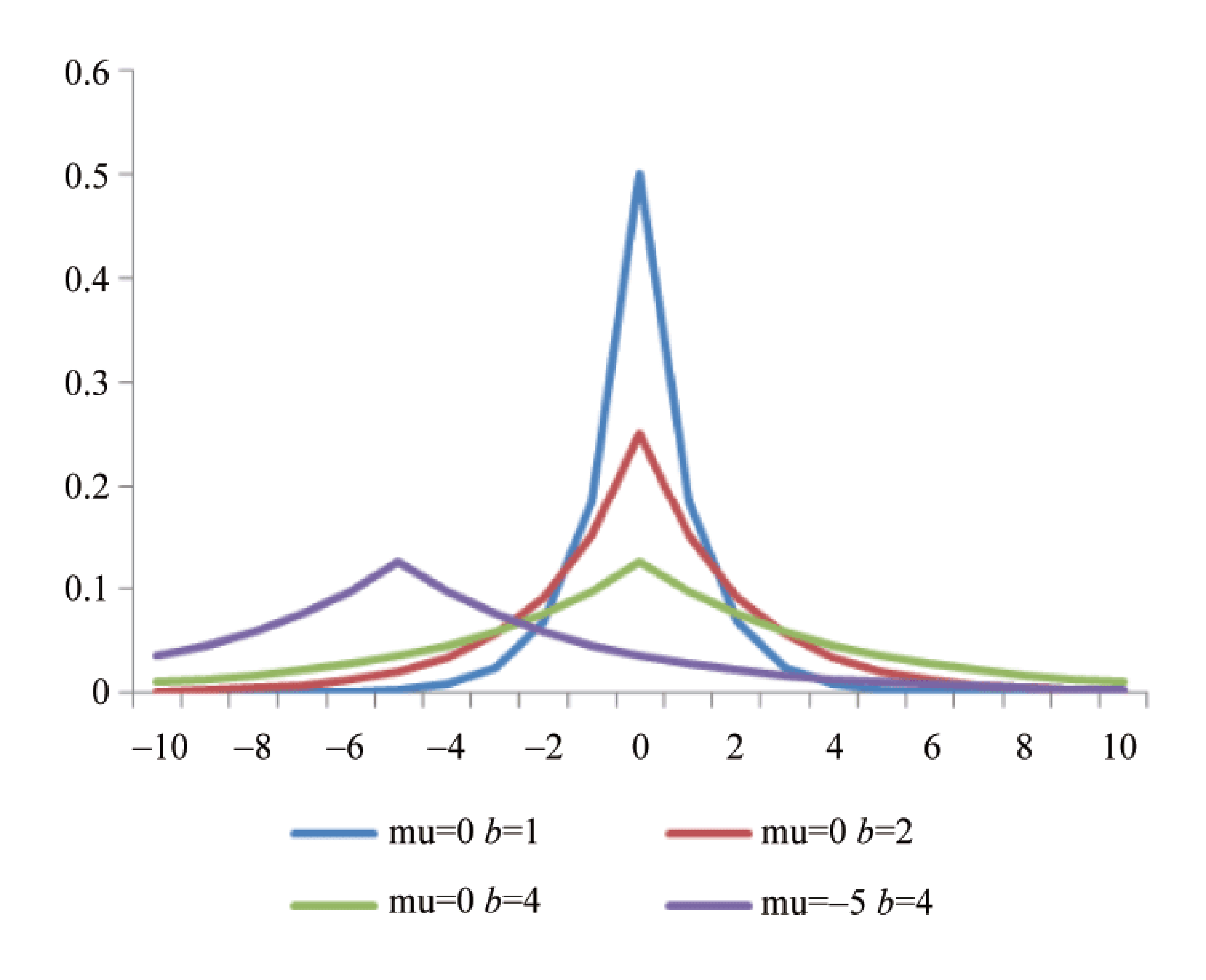


图 7.28　拉普拉斯概率密度函数分布

共建数据隐私新秩序

数据的勘探者看到了数据的价值，那确实是令人激动的价值，这不仅仅是为了一己之私或纯粹的经济利益，他们想通过数据来构建美丽的新世界。

——埃里克·西格尔（Eric Siegel），《大数据预测》

数据使用需要秩序的约束和保护

就像人类掌握了火是迈向文明的第一次伟大实践，人们对数据的大规模使用再一次将文明推进到赛博时代。卡耐基梅隆大学的汤姆·米切尔（Tom Mitchell）教授曾在《科学》杂志上写道："对于定位数据（通过 GPS 对手机追

踪实施定位)的应用可使我们获得各类福利,例如减少交通拥堵、降低污染、控制疾病传播,以及提高公园、公共汽车和救护车等公共设施的利用效率。”我们肯定数据的价值、力量和重要意义,正如前面另外的章节介绍的那样,但我们仍会看到个人隐私信息面临的潜在泄露风险。当下互联网作为基础设施,数据在终端与平台之间、服务与公司之间甚至国家之间几乎能够瞬间转移,数据涉及的用户众多,而数据收集者采用的方式方法透明度很低,数据在流动过程中多方介入缺乏控制力,更不用说数据存在被买卖的情况以及新出现的以数据为商品的交易市场。

亚历克西斯·马德里在《大西洋月刊》中指出:“当今你的简介也许能通过一分钱或者更少的钱来买到,谷歌掌握的你那部分数据价值约20美元,Facebook的是5美元等等,但一个用户对于互联网广告业的价值可能是每年1200美元。”

对于数据信息越来越高的使用价值,除了通过立法明确数据保护相关政策、借助行业标准规范行业行为来维护用户数据的所有权和使用权外,也需要用户与企业的积极互动。不少人在安装一款软件后,都被询问过是否要加入“用户体验改进计划”。然而多数用户并不明白这个所谓的用户体验改进计划意味着什么,即使通读了说明获知该计划的目的是改善服务体验,用户也不清楚自己要为这项计划付出什么样的代价。事实上,**在我们点击确认按钮的刹那,我们就将自己的相关数据发送和反馈给了那些搜集信息的软件公司,这不能算作是与用户的沟通**。而这样的事情早在十几年前就开始了。2003年9月,微软公司针对Windows操作系统推出的办公套装软件Office 2003就启用了一定的算法,把用户在真实环境下使用软件的行为数据连同机器配置等信息记录一并通过互联网发回到公司内部,“用户体验改进计划”这个名字在那时被发明出来,微软公司定义这个项目代号为SQM。

詹森·哈里斯(Jensen Harris)的博客记录(MSDN.COM)曾显示:“自

Office 2003 发布以来，收集的使用片段一共有 13 亿个，每个使用片段都记录了某个固定时间内的所有 SQM 数据，连续 90 天内，仅 Word 就记录了超过 3.5 亿次的命令行点击。”能有那么多用户在真实使用场景下提供行为数据是有些令人吃惊的，对于微软这样大的一家软件公司，当时数据多到有些收不过来。

有些时候，伴随着弹框提示我们也能看到“隐私条款及免责声明”之类的字样，**那些具体的声明信息冗长且难以理解，大多数普通用户根本不会去阅读，这也不是与用户沟通**。企业常常因担心引发谴责，与用户使用不同的语言。为此，在未来如何找到一种简便合适的沟通方式，减少企业与用户之间相互理解的障碍是非常重要的。

埃里克·西格尔博士作为美国预测分析领域的专家，在解析大数据预测工作可能面临的问题时抱有这样的担忧：“各家公司满足于继续神秘地完成用户数据搜集，因为担心引发谴责，但谴责迟早会发生。”他还认为：“最好让人们现在树立意识，选择该分享什么，不该分享什么，以及如何、在哪里进行分享。”

对于用户来说，还以 Facebook 为例进行估算，根据上面提到的 Facebook 掌握每个用户的信息数据每年能带来 5 美元的营收，如果用户愿意出这 5 美元，就可能让 Facebook 放弃输出自己的个人信息，如果用户愿意多出一些，比如 10 美元，就可能得到 Facebook 用隐私保护算法支持的安全服务。换算起来一个月不到一美元的费用，还不到很多地区有线电视费的价格，却对保障我们个人信息隐私有很大益处，特别是给像 Facebook 一样的网络服务提供商传递了一种新的动力，即他们可以在研发和设计网络应用的时候想尽办法多支持、更好地配套不同级别的隐私保护安全服务，向用户提供个人信息保护成本和安全程度的多种选择同样可以获取收益。毕竟使用隐私保护算法也会有一定的技术开销。

当然，我们知道，不论是 Tails 匿名系统，还是 k-匿名、l-多样性算法甚至是差分隐私保护算法，都不能保证隐私的绝对安全。事实上，这个世界上不存在绝对的安全。但是，**隐私并不是一场零和游戏，算法正在用一种更为公平的方式，重塑用户个人与数据公司之间的信任与合作，建立数据使用的新秩序**。

第 8 章　信任的基础——区块链

美国经济学家、诺贝尔经济学奖得主肯尼斯·约瑟夫·阿罗认为:“信任是经济交换的润滑剂,是控制契约的最有效机制,是含蓄的契约,是不容易买到的独特的商品”“世界上很多国家的经济落后都可以通过缺少相互信任来解释”。的确如此,如果人们对银行失去信任,就会发生抢兑;如果人们对股市失去信任,股市就会崩盘;如果人们对国家的未来失去信任,经济危机便会迅速爆发。信任是社会系统和经济系统得以正常运行的基石。越是发达的社会,人们彼此间的信任程度越高。

一般情况下,大家都认为最值得信任的机构一定是政府和银行。银行成为普通百姓储蓄、理财、贷款等金融活动的最主要途径,因为在这些人的心目中,银行是由权威机构(如政府)设立的,甚至可以说,只要国家没有灭亡,这种信任关系就是坚实可靠的。然而,看似坚不可摧的金融巨头也存在破产倒闭的风险。让我们看看 2008 年的美国“次贷危机”。图 8.1 详细展示了美国“次贷危机”的发生过程。

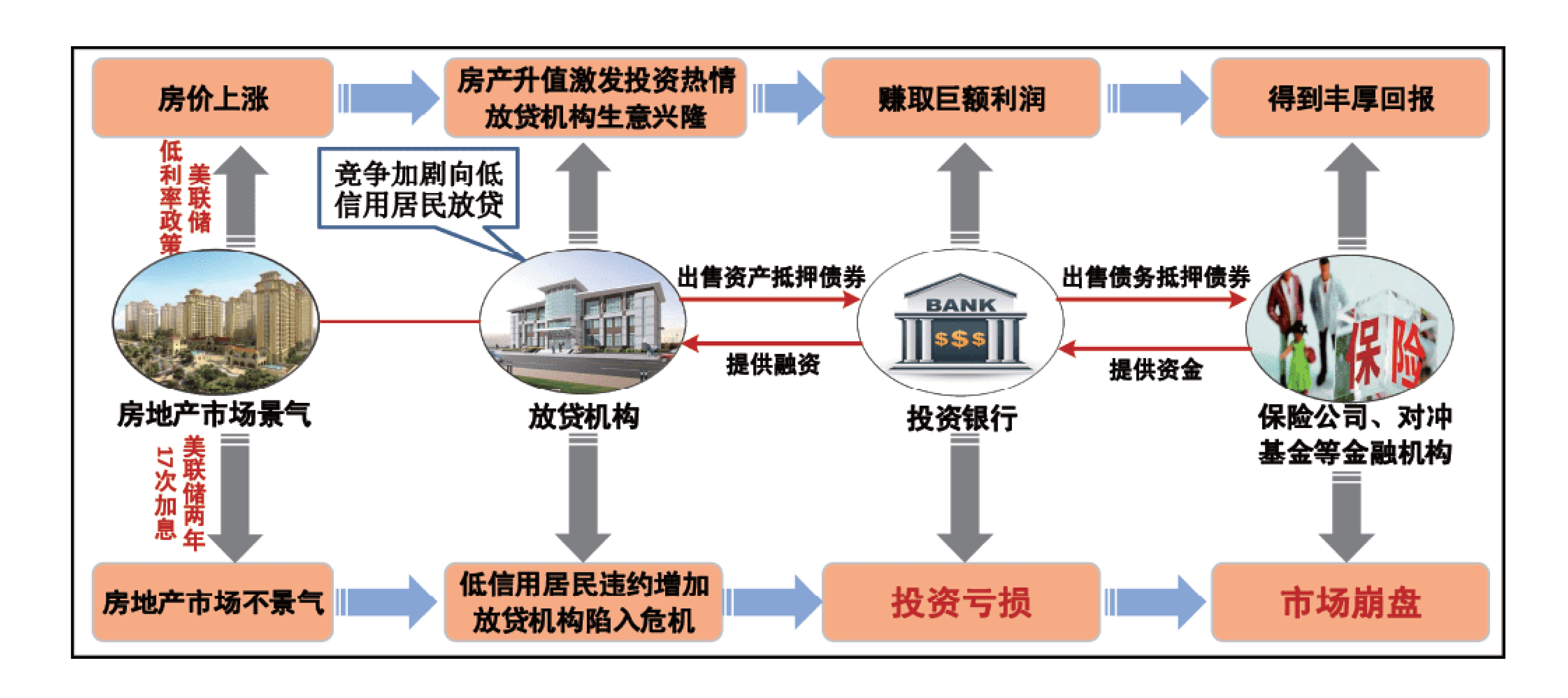


图 8.1　美国“次贷危机”

21 世纪初，美国互联网泡沫破裂，布什政府为了缓解国内经济萧条的局面，于 2000—2006 年通过美联储降低利率、提升减税幅度等相关措施推动了美国房地产市场的发展，从而在一定程度上缓解了美国经济不景气的局面。在这几年时间内，美国房价疯涨，房地产市场一片繁荣，进而推动了放贷机构生意的火爆，但由于竞争加剧，这些放贷机构不得不向低信用的居民发放贷款，并向投资银行出售资产抵押债券从而获得融资，这使得投资银行、保险公司、对冲基金等金融机构赚取了巨额利润。但好景不长，美联储两年内 17 次的加息政策使得房地产市场逐渐萧条，大批次级信用贷款者无法偿还贷款，大量违约现象的出现使得放贷机构陷入危机。这种情况下，投资银行出现巨额亏损，保险公司、对冲基金等金融机构因资金亏损出现大规模倒闭，最终导致市场崩盘，美国“次贷危机”全面爆发。

普通美国人成了这场灾难的受害者，他们一直被他们所信任的银行及金融机构蒙在鼓里，随着他们的财富所剩无几，他们的信任也被逐渐瓦解，最终导致了一场席卷全球的金融危机。所以，无论是市场经济还是计划经济，首先都必须是法制经济、契约经济，也是信用经济。社会信任是推动经济增长

的一个重要因素，是塑造经济社会的先决条件，是决定经济增长方式和企业组织形式的关键因素。建立起符合经济规范的社会信任体系，才有经济的健康可持续发展。一个缺乏信任的社会将加剧交易难度，降低经济运行效率，同时削弱了本国商品的市场竞争力，优质产品难以正常销售，市场鱼龙混杂，群众对任何宣传、广告都不敢轻易接受，从而令经济进入一个恶性循环。

日裔美籍学者福山所著的《信任：社会道德与繁荣的创造》一书深入分析了社会信任程度及其范围对经济组织形态、产业结构、经济运行效率乃至国家竞争力的重大意义。

福山将华人社会（包括中国大陆、香港地区、台湾地区）、意大利、法国、韩国这 4 个低信任度的代表作为一类，在这些国家或地区中，传统文化或宗教崇尚家族主义，导致人与人之间的信任程度仅限于家庭或家族的范围内，对于其他人，特别是陌生人基本上没有信任。另一类则是以日本和德国为代表的以非血缘关系的社团为基础的高信任度国家。在文化和历史上，较少的中央集权历史、家族力量的相对薄弱与中间社会连属关系的强大，促使人们将信任扩展至家族之外。

在经济组织的过程中，低信任度的国家或地区呈现出大量的小型家族企业，但由于社会普遍信任度低及社会资本的缺乏，企业无法随着发展的需要顺畅地迈入现代专业化大型企业阶段，加之家族的财产均分继承制度和排斥外人经营，企业往往富不过三代。为了发展大型企业，政府只能积极介入，扶持家族企业难以涉足的资本密集型加技术密集型大企业，这些大企业又往往处于国家竞争中居于重要地位的产业，在此过程中全社会都不得不承受效率的损失与资源的浪费。而高信任度的国家则往往更顺利地从家族小企业演变为现代专业化管理的大企业。

无论是经济组织还是政治组织形态，即市场经济与民主政治，并不是宏大的冷冰冰的固态，而是社会中每个人行为的放大与集合。一个各怀鬼胎的

集体一定比不上一个团结的、有共同价值观和凝聚力的集体有生产力、有效率和有幸福感,因此社会资本与信任度的多寡直接影响到宏观经济行为的效率。在高信任度的国家可以产生企业集团,企业间以互助的形式和道德互惠的传统创造高绩效;而在低信任度的社会,为了堵住各种可能的漏洞,不得不绞尽脑汁将契约搞得很长,不得不常常求助于正式的法律制度(而不是非正式的道德互惠传统)保护脆弱的、缺乏安全感的群体。这些不信任的累加,致使交易成本无形上升,最后结果是由全社会的成员共同买单。

显然我们都应该同意福山的观点,中国确实是一个低信任度的国家,社会和经济系统存在的漏洞在不时地瓦解人们之间的互信。赛博经济越是飞速发展,就越需要首先解决这些信任问题。以阿里巴巴这样的电商平台为例,它能够让商品在线交易,首先需要建立一个能够消除买卖双方异地交易、陌生人交易可能产生的风险的信任平台,于是支付宝作为第三方支付担保平台就产生了。从 2003 年支付宝首次在淘宝上推出,到 2015 年“双十一”912 亿人民币成交量,再到 2016 年单日进入千亿元,背后却是支付宝推动阿里生态从裂变到聚变的过程。然而支付安全并不能完全保证信任,淘宝平台的信任问题依然在不断地发生,2015 年 1 月至 9 月,支付宝系列共出现 12 个漏洞,平均每个月就有 1.3 个漏洞;在 2017 年的“两会”上,马可波罗瓷砖董事长黄建平表示,在淘宝上有 300 多家马可波罗的店铺销售品牌瓷砖,但只有两家店铺获得授权,其余都是侵权的假冒伪劣商品。虽然阿里巴巴董事长马云一直强调要“像治理酒驾一样治理假货”,但这并不能避免人们对赛博经济下第三方平台的安全和可靠性的怀疑。

无论是政府管辖的中央银行,还是企业设立的第三方平台,它们都是高度依赖中心化处理的金融机构。中心化处理效率高,成本可控,但是一旦中心出现问题,系统将迅速崩溃,特别是当人们对中心产生信任危机时,中心化的方案就会遇到很大困难。

经历过一系列危机后，人们可能都会有这样的疑问：我存入银行的钱现在在哪？次贷危机的悲剧会不会发生在我身上？有一天企业如果倒闭了，我该向谁要这笔钱？这些对第三方平台及银行的不信任也在促使人们思考，是否存在一种不依赖中心化的信任机制，可以保证上述经济运行的安全可靠。

诞生于“次贷危机”之后的区块链就是运用去中心化思想的共识机制，实现信息的透明化与公开化，进而令重建一个公开透明的技术监管和可信的金融系统成为可能。以“次贷危机”为例，其发生的诱因是住房抵押贷款证券价值大为贬值，人们原本以为值钱的东西突然不值钱了。然而，人们最初并没有对某些商品的价值达成共识，而是将这些商品价值的评估依赖于一些外部条件，比如第三方的评估。反之，倘若所有人都对某一商品的价值达成共识，那么就不存在商品的大幅贬值问题。区块链就为我们提供了一种达成这种共识的基础技术。

拜占庭将军问题

这座城市的中央计算机告诉你的？R2D2，你不该相信一台陌生的计算机！

——C3PO，星球大战中天行者阿纳金制造的机器人

如何才能让我信任你

自古以来，人类社会里许多规则的设立，都是为了创造一个信任的环境，解决信任问题，让更多人从中受益，大到一个国家的法律法规，小到一个公司

的制度章程。然而，尽管人类在不断完善自己的社会信任体系，依然难以避免那些在特殊情况下会发生的信任危机。在拜占庭时代，拜占庭将军问题给了人们一个启发，或许我们可以利用逻辑推理和数学算法去解决这类基于**在一个互不信任的分布式群体中达成信息共识**的难题。下面我们仔细分析拜占庭将军问题①，并探寻如何通过算法来解决这个问题。

在很久以前的拜占庭时代，有一个繁荣富强、兵强马壮的国度——拜占庭，它的帝王把整个国家治理得井井有条，但也会时常抢掠周围几个城邦的财富（为方便描述，我们假设周围有 4 个城邦）。而作为各自城邦的军事领袖，将军 A、B、C、D 实在忍受不了拜占庭帝国的欺凌，但在庞大的拜占庭面前，他们各自的实力实在太过渺小，若他们单独进攻，犹如“以卵击石”。经过实力对比，他们各自发现只有超过半数的将军（即至少 3 位将军）在同一时间进攻拜占庭才能取得最终的胜利，否则他们各自国家的财富也会被“作壁上观”的城邦抢光，所以他们希望能够联手攻打拜占庭帝国。

但是问题来了，由于历史遗留原因，4 位城邦的将军 A、B、C、D 彼此都不信任，他们可能出尔反尔，答应了出兵却又临时变卦。同时，他们没有一个盟主来统一规划他们的出兵计划，而且出于防止被劫持、谋杀等安全因素的考虑，他们也不可能聚集到一处来商讨进攻拜占庭的时间。面对这种情况，4 位将军只能通过各自的信使来传达进攻时间信息。例如，将军 A 让信使向其他三位将军传达“将军 A 希望进攻时间是上午 9:00”，将军 B、将军 C 和将军 D 也在同一时间以同样的方式分别向其他三位将军传达 11:00、13:00、18:00 进攻拜占庭的消息。为了达成共识，每位将军需要根据收到的消息再

① 拜占庭将军问题是图灵奖获得者 Lesile Lamport 于 1982 年首先提出的，这个问题的命名是 Lamport 受 Dijkstra 哲学家吃饭问题的启发而杜撰的，他认为让问题有历史感有助于其被公众广泛认知，实际效果却并不理想，直到比特币诞生后，这个问题才被公众了解。

次选择一个进攻时间,并在同一时间再次发送给另外三位将军。由于每一位将军有 4 种选择,所以 4 位将军选择的进攻时间共有 256 种的可能情况。在这些情况中,只有至少 3 位将军同时达成一致共识时才能出兵拜占庭,而这样的结果有 64 种,也就是 4 位将军每次仅有 1/4 的概率达成共识,而且随着将军数量的增多,这种共识的达成会变得更加渺茫。不仅如此,4 位将军互不信任,他们之中可能有一人背叛他人(称为"叛徒"),违背之前达成的攻占计划,"作壁上观"等待他人兵败拜占庭,进而"坐收渔翁之利"。在这种情形下,如何让 4 位将军更加容易地达成共识成为一个难题,这就是所谓的"**拜占庭将军问题**"(**图 8.2**)。

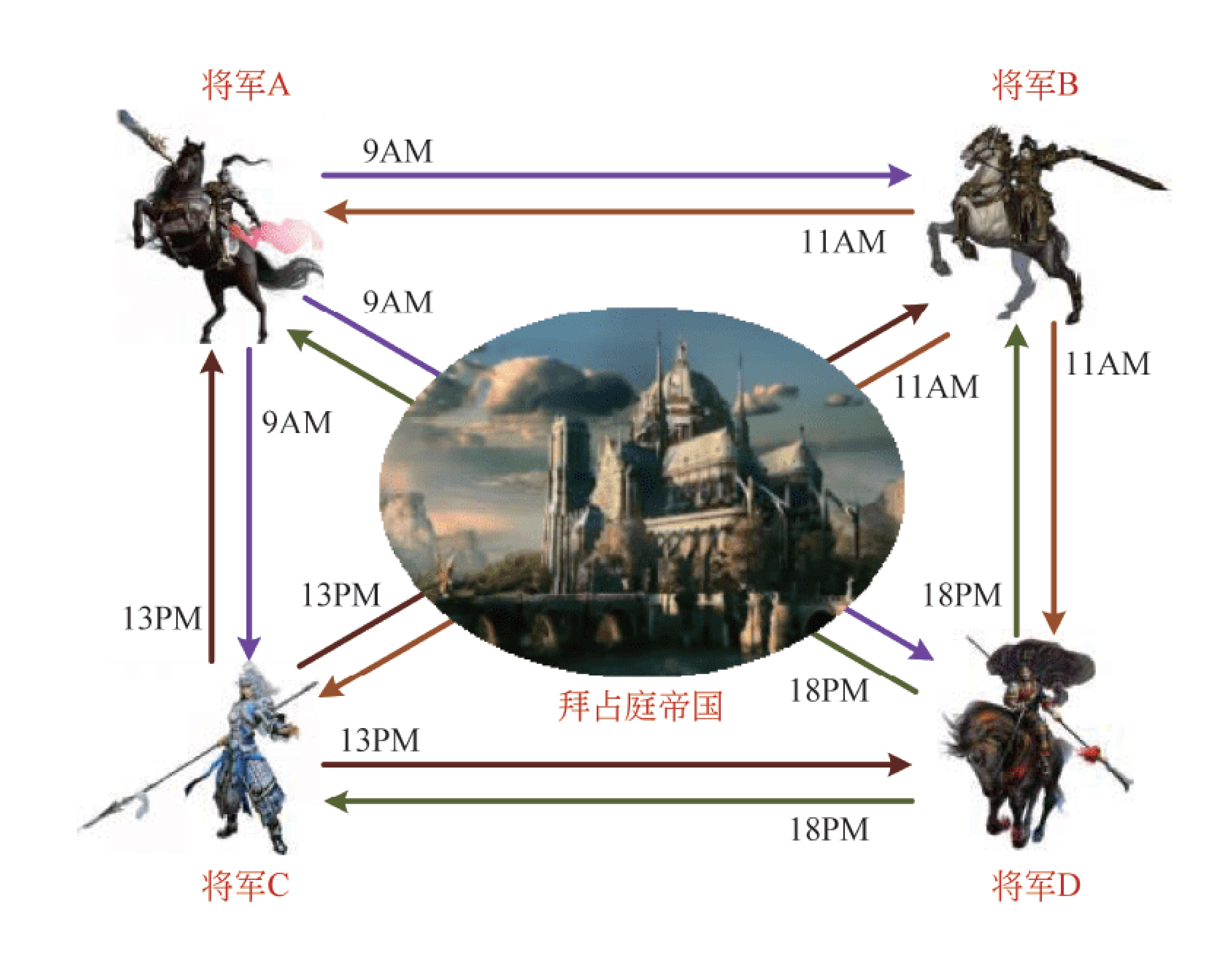


图 8.2　拜占庭将军问题示意图

在讨论"拜占庭将军问题"时,假设消息传递过程准确无误,即不存在消息被丢弃、消息遭篡改等情况,我们主要关注众多将军中是否有叛徒。在上面的例子中,出于不信任的关系,4 位将军可能存在叛徒,会"临时变卦"或"假传圣旨",这也给他们达成共识增加了难度。而在中国古代,破坏彼此共

识的事例不胜枚举。

春秋战国时期，公元前630年，因郑国从属晋国的敌人楚国，且之前没有礼待晋文公，晋国联合盟友秦国围攻郑国，在这千钧一发之际，秦穆公在烛之武的劝说下却选择退兵，因为秦国害怕晋国强大之后会危及自己的疆土，这就是历史上有名的“烛之武退秦师”(图8.3)。



图8.3 烛之武退秦师

安史之乱之后，唐朝一片萧条，公元765年，吐蕃联合回纥、吐谷浑以及山贼等30万军队攻打长安。唐朝大将郭子仪只身前往回纥阵营劝说回纥首领药葛罗，在郭子仪的劝说下，回纥与唐朝结盟，并联合出兵吐蕃，给吐蕃以沉重的打击，从而帮助唐朝避免一场浩劫。

秦国和回纥破坏了曾经达成的共识，导致了原来的同盟方遭受重大失败。因此让共识达成一致且不被破坏，成为“拜占庭将军问题”的核心内容。

信任这类问题是有其内在逻辑的，首先叛徒希望达到以下目的：(1)欺骗某些将军，促成“在某个时间发兵”的虚假共识；(2)混淆某些将军，扰乱其他将军达成共识。所以，各位将军必须有一套抗干扰的机制，才能使所有忠诚的将军达成何时出兵拜占庭的共识，并且不受少数叛徒欺骗和混淆的影响。

经过推理，我们可以知道当叛徒的数量少于 1/3 时，“拜占庭将军问题”可以解决[①]：举例来说，假设只有将军 C 一个人是叛徒。不妨站在将军 A 的角度来分析，每次将军 A 收到的三条消息里有两条消息是来自忠诚的将军 B 和将军 D，而这两条消息是正确的（可靠的）；另外一条消息来自叛徒将军 C，而这条消息是错误的（不可靠的）。即将军 A 收到的三条消息里，有两条消息是正确的，一条消息是错误的。

如图 8.4 所示，假如忠诚的将军 B 和将军 D 都想“发兵”，而叛徒将军 C 见机不妙，想破坏这个共识，所以故意说“撤退”。但将军 A 收到的“发兵”消息数量要多于“撤退”消息数量，所以他也决定“发兵”。这样，即使叛徒将军 C 最终选择了“撤退”，因为有将军 A、将军 B 和将军 D 攻占拜占庭，这样也会取得最终的胜利，“拜占庭将军问题”得以解决。



图 8.4　叛徒数量少于 1/3

叛徒数量等于或多于 1/3 时，“拜占庭将军问题”不能解决：假设将军 C 和将军 D 两个人都是叛徒。不妨站在将军 A 的角度来分析，每次他收到的三条消息里有一条是来自忠诚的将军 B，而这条消息是正确的（可靠的）；另外两条消息则来自叛徒将军 C 和将军 D，而这两条消息是错误的（不可靠的）。即将军 A 收到的三条消息里，有一条消息是正确的，两位两条消息是错误的。

如图 8.5 所示，假如忠诚的将军 B 想“发兵”，而叛徒将军 C 和将军 D 都故意想“促成”这个共识的达成，因为如果仅有将军 A 和将军 B“发兵”的话，一定兵败拜占庭，这样他们两个城邦的财富就可被叛徒将军 C 和将军 D 抢

① 细心的读者可能觉得叛徒数量不多于 1/2 就可以，这个问题比较复杂，一个简单的例子是当三个将军中有一个叛徒时，另外两个人是无法判断谁是叛徒的，从而无法达成一致。感兴趣的读者可以参考 1982 年 Lamport 的文章 *The Byzantine Generals Problem*。

走，所以两个叛徒都谎称自己想“发兵”。这样一来，将军A会收到三条“发兵”消息，他会义无反顾地“发兵”。而最终结果却只有忠诚的将军A和将军B攻打拜占庭，他们不仅战败，而且自己的财富也被叛徒将军C和将军D抢劫。在这个例子中，因为叛徒数量超过一定界限而使得共识问题遭到破坏，在这种情况下，“拜占庭将军问题”不可解决。



图 8.5　叛徒数量多于 1/3

从上面的分析可知，在同步通信环境中，叛徒个数小于将军总数的1/3时，将军们可以达成一致命令。如果同步通信是可认证、防篡改的，任意多叛徒都可以有解决方案。而如果异步通信时，只要有一个叛徒存在，“拜占庭将军问题”便无解。

“拜占庭将军问题”提出了在一个互不信任的分布式网络系统中达成可信共识的问题，然而我们的社会跟我们的网络系统一样，远比这复杂。如果把这个问题放到我们的网络系统中，那么叛徒就代表网络异常节点，当异常节点较少或出现的概率较低时，关于“拜占庭将军问题”的讨论才有意义。在计算机领域，管理员错误配置、误操作、网络节点遭受黑客或病毒攻击、系统后门、漏洞等原因都可以导致网络节点异常，进而使得整个网络系统不可信，这也给网络系统中众多节点达成共识增加了难度，所以我们需要进一步解决在不可信的网络环境下如何达成共识。

拜占庭将军协议及容错系统

你不可能通过挑战既定事实来改变一件事情，除非建立一个全新的技术

并废弃现有机制。

——R.Buckminster Fuller，美国建筑师、作家、设计师、发明家

为了解决拜占庭将军问题，Lamport① 和 Pease② 等学者均提出了各自的解决方案，这些方案有一个共同点，那就是使用了递归算法，普通读者理解起来并不容易。清华大学姚期智先生(2000 年图灵奖获得者)的学生王君行最近提出了一种新型网络协议③，用简单的循环替代了复杂的递归运算，算法简单直观。值得一提的是，王君行提出这种解决方案的时候还是一名大一新生，让人不禁感慨“自古英雄出少年”。下面讲述一下这个精巧算法。

假设 n 个拜占庭将军($G_0, G_1, \cdots, G_{n-1}$)中有 t($n>3t$)个叛徒(是的，叛徒数量必须已知，Lamport 的算法也一样)，每位将军的消息用 r 表示。集合 A 为集合$\{1,2,\cdots,n-1\}$的子集，其数量为忠诚将军的数量，即$|A|=n-t$。一开始，G_0将消息 r_0广播至所有将军，对于每个集合 A，其中的所有将军将收到来自 G_0的消息 r_0以自身消息 r_i广播至所有将军，每个将军将收到的多数

```
Commander G0 sends r0 to all.
for A ⊆ {1, 2, ···, n − 1}, |A| = n − t do
    for i ∈ A do
        Gi sends ri to all;
    end for
    for j ∈ {1, 2, ···, n − 1} do
        rj ← majority(Rj←A)
    end for
end for
Each Gi takes ri as his answer.
```

图 8.6　简单的拜占庭将军协议算法

① Lamport L，Shostak R，Pease M. The Byzantine Generals Problem[J]. ACM Transactions on Programming Languages and Systems (TOPLAS)，1982，4(3)：382-401.

② Pease M，Shostak R，Lamport L. Reaching Agreement in the Presence of Faults[J]. Journal of the ACM (JACM)，1980，27(2)：228-234.

③ Wang J. A Simple Byzantine Generals Protocol[J]. Journal of Combinatorial Optimization，2014，27(3)：541-544.

将军发送来的消息视为自身的最终消息。这样，当算法执行完成后，n 个拜占庭将军就 G_0 发送的消息均能达成共识。

为了便于理解上述算法，我们举例说明，假设在一个网络中存在 7 个节点：$G_0, G_1, \cdots, G_6$，且 G_0 向其他 6 个节点发送消息。此时的目标是让所有的 7 个节点能够达成共识，即收到的消息是一致的。需要说明的是，网络中可能存在恶意节点。当存在 2 个恶意节点时，不妨假设 G_1、G_2 为恶意节点，G_0 利用广播技术向其他 6 个节点发送消息“1”（或者说 G_0 希望消息“1”成为大家的共同知识）。针对该消息，从剩余 $7-1=6$ 个节点的集合中选取包含 $7-2=5$ 个节点的子集，这样的子集共有 6 个，如图 8.7 所示。对于集合 A 的任意一种情况，可靠节点的数量总是大于恶意节点的数量，所以每次循环的结果都不会受到恶意节点的干扰，直至集合 A 的所有情况都被执行完毕，最终达到所有节点的共识。

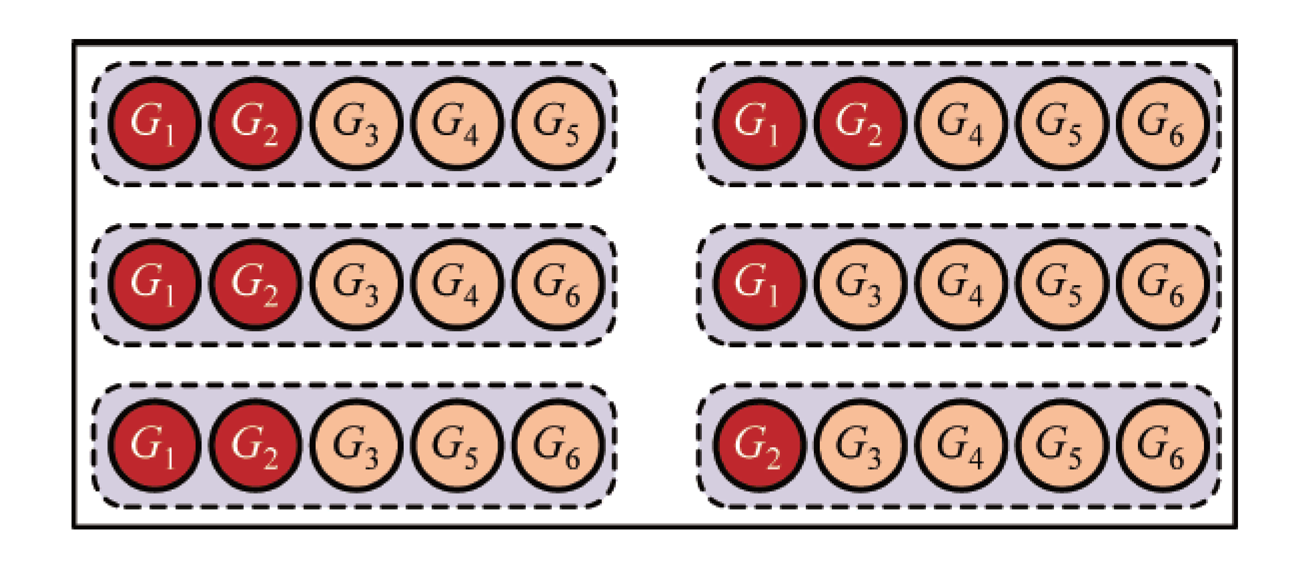


图 8.7　集合 A 的所有情况

需要指出的是，算法巧妙利用了单调性，也就是说一旦忠诚的节点达成共识，该共识就会一直持续下去。如果 G_0 忠诚，那么共识一直持续（图 8.7 的情况）。如果 G_0 不忠诚，算法会遍历所有忠诚节点数量的子集，则也一定能取到一个忠诚节点的完整集合，这时大家可以达成共识，然后持续到算法结束。要注意，虽然取到忠诚节点的全集，但是由于 G_0 不忠诚，他会告诉不同的节点不同的命令，但是由于忠诚的节点会忠实的交换收到的命令，所以

最终仍然可以达成共识。

需要注意的是，在拜占庭将军问题中，只有叛徒数量少于三分之一的将军总数时，该问题才可解。具体证明可以参考 Lamport 的论文。

2008 年图灵奖得主芭芭拉·利斯科夫[1] 1999 年在论文 *Practical Byzantine Fault Tolerance*[2] 中提出了一种实用的拜占庭容错系统。具体机制如下：

系统中的节点通过选举的方式选择出主节点(类似于领导选举)，图 8.8 中节点 0 为节点 1、2、3 的主节点，其中节点 3 为恶意节点。当客户端向这些节点发送消息时，首先将请求消息发送至主节点 0(request 阶段)，节点 0 将来自客户端的请求消息发送至节点 1、2、3(pre-prepare 阶段)，这三个节点在收到来自主节点 0 的请求后，会通过广播的形式告诉彼此该消息是什么(prepare 阶段)，在收到别人的“告知”后，每个节点根据多数节点的“告知”决定出请求消息的内容，并执行该请求(commit 阶段)，最后将执行结果反馈给客户端(reply 阶段)。虽然系统中存在恶意节点 3，但这并不会影响整个系统对客户端请求的执行。

当然，并不是所有选举的主节点都是可信的，例如图 8.8 中，如果节点 0 为恶意节点，而其他节点为可信节点时，各个节点会根据由主节点 0 下发的消息进行判断(也是通过广播的形式获得其他节点收到的消息)，如果主节点 0 不可信，则该系统将进行重新选举主节点。这种检查与选举机制在一定程度上保证了整个系统达成的每个共识都是安全可信的。

---

① **芭芭拉·利斯科夫**(1939 年—)，本名 Barbara Jane Huberman。美国计算机科学家，2008 年图灵奖得主，2004 年约翰·冯诺依曼奖得主。美国工程院院士，美国艺术与科学院院士，ACM 会士，现任麻省理工学院电子电气与计算机科学系教授。她是美国第一个计算机科学女博士。导师为 1971 年图灵奖得主约翰·麦卡锡。

② Castro M, Liskov B. *Practical Byzantine fault tolerance*[C]//OSDI. 1999, 99: 173-186.

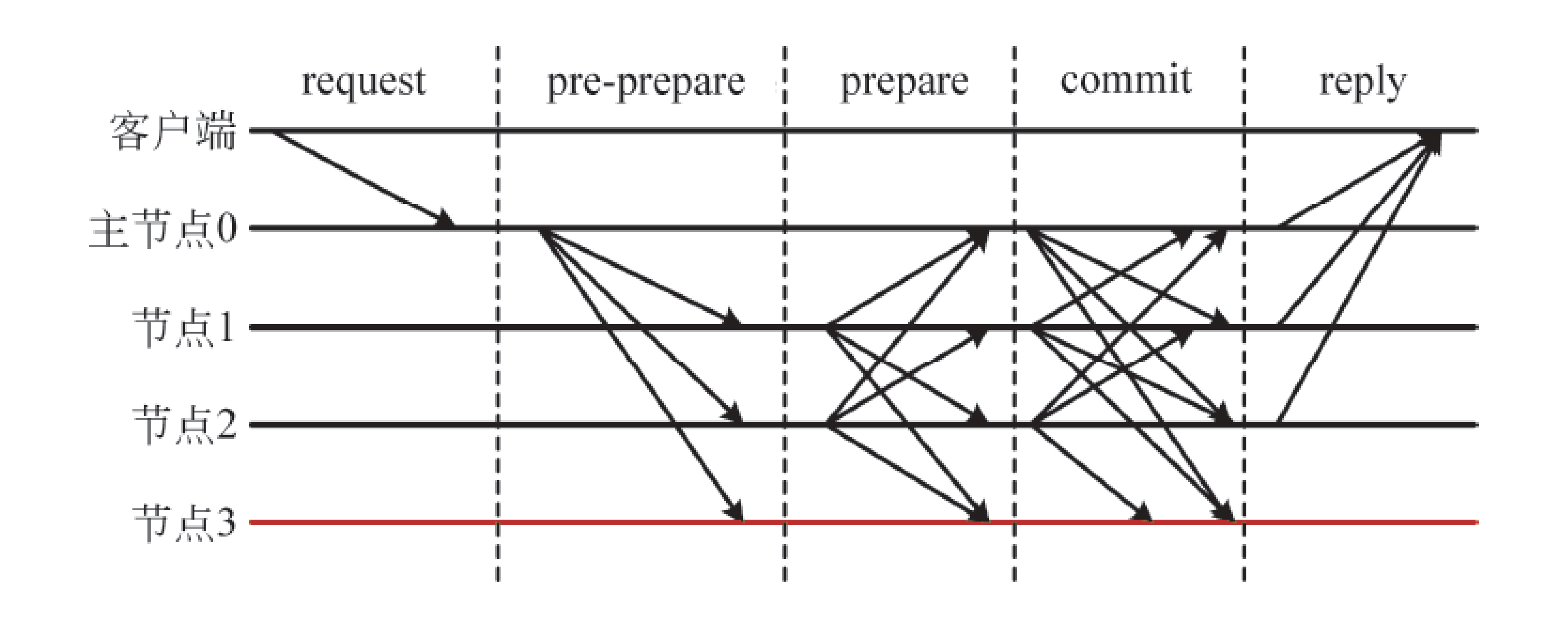


图 8.8　实用的拜占庭容错系统

共识机制的先驱：P2P

你想用卖糖水来度过余生，还是想要一个机会来改变世界？

——1983 年史蒂夫·乔布斯对当时的百事可乐总裁约翰·斯卡利说

拜占庭将军协议和网络容错技术的发展解决了信息共识问题，这也令P2P 通信中的基本问题——可靠通信问题得到解决。P2P 网络让人们实现了资源的共享，作为 P2P 技术衍生物的区块链，融合了现有技术，它从诞生到现在，能够一直稳定运行，与其采用的分布式网络技术——P2P 技术不无关系。P2P 技术是 peer-to-peer（对等网络）的简称，是一种计算机间直接通信而无须借助中心化设备来共享计算机资源和服务的分布式技术。它没有真正意义上的中心，各个网络节点直接通信、相互协作以达到资源共享、服务共享的目的。其实，P2P 技术对于我们来说并不陌生，PPLive、PPS 网络电视、QQLive 以及曾经饱受争议的“快播”都是 P2P 技术的典型应用。

彪炳千古的 P2P

现代生活中，P2P 虽不是一项新的技术，但在 20 世纪 90 年代末，它却满足了数以千万的用户对网络资源下载的需求，以至于迅速风靡全球。同时，几场官司也使得 P2P 成为互联网界绯闻缠身的“名人”。

1998 年，美国东北大学的一年级新生，18 岁的肖恩·范宁（Shawn Fanning）为解决舍友“如何在网上寻找音乐”的问题，开发出了 Napster 系统，成为 P2P 技术应用的先锋。Napster 将所有音乐文件的地址存放到一个集中的服务器中，并对用户提供音乐检索功能，这样用户就能方便地找到所要的音乐文件。Napster 让无数音乐爱好者趋之若鹜，被认为是第一个真正有影响力的 P2P 软件，最高峰时有 8000 万注册用户。1999 年，Napster 公司成立，它能够提供音乐作品上传、检索和下载的功能，一时之间蓬勃发展。但好景不长，美国唱片业协会（RIAA）起诉 Napster 歌曲侵权，败诉后的 Napster 于 2002 年宣告破产。

一个 Napster 倒下，却有更多的 Napster 站起来了：基于 P2P 技术的 Gnutella、Morpheus 和 KaZaA 等公司立即填补了 Napster 所留下的市场空白并加以发扬光大。相比于 Napster，这些公司进行了改进与提升，在避免版权控诉的情况下，使得 P2P 技术的应用更加广泛。首先，这些公司接受了 Napster 的教训，更新了技术，避开了采用集中服务器存储用户文件信息（如音乐文件）的做法，所有的信息全部放在用户的计算机上，它们只提供 P2P 技术，至于用户做什么，与它们没有关系，从而很好地躲过了唱片公司对音乐版权的诉讼，也为 P2P 技术的应用扫清了最大的障碍；其次，这些公司允许用户共享任何类型的文件，而不仅仅是 MP3 类型的音乐文件，扩大了 P2P 技术的应用范围，有力地推动了 P2P 技术的发展。

P2P 技术除了应用于文件共享领域，还在音频通话、视频直播和视频点播等流媒体领域得到了广泛应用，网络即时语音沟通工具 Skype 利用 P2P 技术实现语音服务，CoolStreaming、PPLive 等利用 P2P 技术实现视频播放，Youtube、Youku 利用 P2P 技术加速，提升了用户视频播放体验。

P2P 真正解决了信任难题吗

P2P 的成功并不是一蹴而就的，它的发展历经好几代，新发现的问题不断由下一代来解决，这才造就了 P2P 技术如今的成功。即便如此，P2P 网络中的信任问题仍然没有被完全解决。

以 Napster 为代表的第一代 P2P 系统，主要采用集中索引的方式来处理网络资源的共识问题。中央目录服务器为各个网络客户端提供资源（如音频、视频）检索服务，并将达成的共识（如资源的位置）返回给 P2P 节点，进而节点间直接进行通信。这里，各个客户端主要依赖中央目录服务器来达成某种资源共识。但目录服务器的性能成为这种 P2P 系统的瓶颈。同时，第一代 P2P 系统通信的安全、可靠性必须依赖处于系统中心的目录服务器，这也是第一代 P2P 系统在共识机制上面临的主要问题。

下一代 P2P 系统吸取了经验教训，以 Gnutella、KaZaA 等为代表的第二代 P2P 系统是一种完全无中心的分布式网络，所有的查询和响应都在分布式的 P2P 节点之间完成。用户之间分享各自的软件、媒体等资源不再需要经过中心服务器，取而代之的是以广播的方式散发查询消息给网络中的其他用户，具有较好的容错性。但第二代 P2P 系统在解决了第一代的中心问题的基础上引入了另外一个问题，那就是广播流量的问题。在资源共识达成过程中，需要较高的带宽需求以满足查询请求在网络中的广泛传播，这就给通信、资源共享带来了时间、空间上的限制。在资源的传播过程中，虽然避免了

中心化服务器的安全问题，却没有改变第一代 P2P 系统信息在各个端点处可能被仿冒、篡改等安全问题。

为解决这些问题，第三代 P2P 系统采用分布式哈希表技术[①]来快速查询网络中的文件。分布式哈希表将资源的位置分布式存储在整个网络上，采用关键字和对应的值一一对应的表格方式存储，它可以提供与哈希表一样的快速查询服务。

分布式哈希表允许大量的用户参与哈希表的维护，小部分用户加入和退出对整个哈希表整体影响不大。第三代 P2P 系统以 Chord、CAN、Tapestry 等为代表，同时具备前两代 P2P 结构的高效性和容错性优点。它利用位于主干位置的超级节点，而不是像第一代那样只依赖于中心化的服务器，主干位置的节点不再是单个，而是多个节点共同维护整个体系结构，各个 P2P 节点通过和超级节点交互以达成资源共识（如获得文件信息）并进行文件传输，这种结构对现有的网络具有更好的适应性。这种网络结构能够达成资源共识，但必须要求每个节点都是安全可信的，如果其中一个节点传播错误信息，就会使得整个网络的资源共识无法达成。

因此，如果 P2P 技术缺乏对节点及消息的可靠性考虑，在去中心化结构下无法解决形如消息仿冒、篡改等安全问题，就不能实现消息溯源的功能，而这恰恰限制了跟货币、支付等敏感行业相关的 P2P 技术的发展。

一个不安全的 P2P 网络是无法获取人们的信任的。P2P 的未来将何去何从呢？就在 P2P 技术迷茫之际，比特币（Bitcoin）诞生了。它是一个完全的去中心化、去信任的分布式电子账簿系统，在其基础之上的交易不再需要

---

① 哈希，英文 Hash，就是把任意长度的输入通过哈希算法，变换成固定长度的输出，该输出就是哈希值，也称为散列值。这种转换是一种压缩映射，也就是，散列值的空间通常远小于输入的空间，不同的输入可能会散列成相同的输出，所以不可能从散列值来唯一的确定输入值。分布式哈希指的是哈希值保存在网络中的不同节点，从而可以进行分布式的查找。

可信第三方来保证,人与人之间的信任由密码学的机制来保证,从而更加安全。可以说,比特币用一种创新的方式解决了拜占庭将军问题。

区块链的鼻祖:比特币

我们需要从比特币革命中借鉴经验,但比特币自身还不够完美。

——比尔·盖茨,微软公司创始人

当下,比特币在计算机及金融界越来越炙手可热,而如此火爆的比特币年纪却还不满10岁。

2008年11月,中本聪发表了一篇名为《比特币:一种点对点的电子货币系统》(Bitcoin: A Peer-to-Peer Electronic Cash System)的论文,首次将比特币技术带入人们的视野。次年1月,首个比特币挖矿程序发布,世界上最早的50个比特币诞生。而当前比特币的规模已经空前庞大,市值也已经超过了900亿美元,那么是谁缔造了如此庞大的比特币帝国呢?

“中本聪”首次出现是在2008年11月1日,这一天,“他”在一个密码学网站metzdowd.com的邮件列表中发表了比特币的论文,首次提出了比特币的概念;2009年,全球首款比特币算法软件出现,随后的短短几年时间里,比特币价格疯狂增长,也令其他的数字货币黯然失色。而创始人“中本聪”的身份至今都是一个谜(图8.9)。

2013年5月,计算机科学家泰德·纳尔逊(Ted Nelson)称,日本京都大学教授望月新一就是中本聪。从他的学术成就来看,比特币完全有可能出自他的手。然而,这个观点也引来一片质疑。而望月新一本人也并没有回应他到底是不是中本聪。



图 8.9　中本聪是谁

2014 年 3 月，美国《新闻周刊》一篇报道声称已经找到中本聪，而且与文章作者进行了面谈。据称，中本聪现年 64 岁，又名多利安，日籍美国人，隐居在洛杉矶圣贝纳迪诺山脚下的一座房子里，而且他的真实姓名就是中本聪。当被问及比特币的问题时，他说："我不再参与相关事宜，不能再讨论该问题。我已经把它交给其他人，现在由他们负责，与我不再有任何关系。"令人大跌眼镜的是，在这篇报道发布 3 天后，多利安却又否定了自己是中本聪，而神秘的中本聪也在网上说"多力安不是我"。

2015 年 12 月，美国《连线》和 Gizmodo 网站相继发表文章，认定澳大利亚商人兼学者克雷格·史蒂芬·怀特(Craig Steven Wright)就是中本聪。Gizmodo 调查发现，在比特币出现之前，怀特就曾在电子邮件中讨论了与比特币有关的工作。

由于中本聪掌握着最早产生的 100 万个比特币，因此世人对中本聪到底是谁一直有着巨大的好奇心。到底怀特是不是中本聪，没有人知道，或许中本聪是谁将成为永远的秘密。然而不管怎样，相对于出于其手的比特币来说，他的身份似乎已经不再重要。也有人说，中本聪已经不可能出现了，因为他的出现势必对比特币的市值带来巨大冲击，这种情况下，掌握大量比特币的地下产业不会允许他再次出现。中本聪最明智的做法也许只能是永远隐

藏自己的身份。

从次贷危机到比特币帝国

人们对电子货币的研究比较早，但是比特币却直到2008年才诞生。2008年9月，美国次贷危机全面爆发，这不得不迫使人们对现有的货币金融体系产生怀疑，比特币就是在这样的背景之下诞生的。2009年比特币诞生，中本聪在其中写道："The Times 03/Jan/2009 Chancellor on brink of second bailout for banks"（此处用原文以示敬意），带着对旧体系的嘲讽，比特币真正诞生了。

2008年爆发的美国次贷危机，对于对华尔街和政府的不信任，人们开始寻找一种能够完全独立于政治力量和金融大鳄的电子货币和支付方式；2008年11月1日，中本聪发表题为《比特币：一种点对点式的电子货币系统》的论文，奠定了比特币的理论基础。

2009年1月3日，首个比特币挖矿程序正式发布，中本聪第一个运行该程序，获得世界上最早的一批50个比特币，被称为上帝区块。

2010年5月22日，来自于佛罗里达的程序员Laszlo Hanyecz用10 000个比特币换来了两块比萨；5月22日这一天也被确定为"比特币比萨日"；7月16日，经过为期5天的10倍暴涨，比特币价格从0.008美元升值0.08美元；11月6日，比特币经济总值超过100万美元，每个比特币兑价格达到0.5美元；11月28日，"维基解密"事件发生，比特币帮助维基解密及其创始人阿桑奇度过危机。

2011年2月9日，比特币价格首次达到1美元，与美元等价。此后几个月，比特币先后与英镑等多个国家货币的兑换交易平台开张；3月6日，比特币全网计算速度达900GH/s，但很快又下跌了40%，显卡挖矿开始流行；6

月19日，黑客从感染木马的计算机上盗用了用户的MT.Gox证书，6万个用户数据被泄露，导致875万美元的账户受影响，此后数月多家平台被黑，Bitconinica平台因两次遭受攻击，最终停了服务。

2012年3月1日，服务器超级管理密码泄漏，价值228 845美元的46 703个比特币失窃，黑客是比特币世界挥之不去的噩梦；9月27日，比特币基金创立，此时一个比特币为12.46美元；11月28日，区块供应量首次减半调整，从之前每10分钟50个递减至25个，同时比特币发行量已经达到发行总量2100万的一半，此时一个比特币为12.4美元；12月6日，世界首家官方认可的比特币交易所——法国比特币中央交易所诞生，此时一个比特币为13.69美元。

2013年3月28日，比特币总市值超过10亿美元，一个比特币均价为92美元；4月10日，比特币价格创下历史新高，一个比特币为266美元；4月20日，四川芦山地震当天，李笑来在bitcoin官网发起对灾区的比特币捐赠，中国壹基金此后宣称共计收到捐赠比特币233个，市值22万人民币，此时比特币价格理性回落，一个比特币为121美元；8月9日，德国成为全球首个认可比特币的国家；10月29日，加拿大启用世界首台比特币自动提款机；11月29日，比特币价格达到历史上的新高，一个比特币为1242美元，而当天的黄金价格是每盎司1240美元。至此，比特币实现了从0.008美元到1242美元三年狂涨千万倍的伟大壮举。

从2014年到2017年，越来越多的人疯狂追求挖矿，迷恋比特币，促成了比特币的成功。那么比特币作为一种电子货币，是如何解决信任、共识问题从而被大家认可和追捧的呢？

“比特县”中的比特币

截止到2017年3月3日，共有16 195 638个比特币已被挖出，根据比特币系统所使用的技术机制，系统设定的上限是2100万个比特币，预计在2040年所有的比特币将全部被挖出。

话说在拜占庭帝国，有一个经济非常发达的“全国十强县”比特县。但有一个令人不解的问题：比特县竟然没有一家银行！原因就在于该县的百姓全部利用一种叫做“比特币”的电子货币进行交易，“比特币”成为这个县代名词，也是比特县名字的由来原因。

很久以前，比特县的百姓都是用“以物易物”的方式进行交易，比如老王家用一袋大米换老赵家一只母鸡，李大婶用一罐蜂蜜换王阿姨半筐冬枣，但这种交易方式实在太不方便，最终被以黄金为代表的实物货币所取代。这时，人们就可以从家里拿出一克黄金购买与之等价的商品，如老王用一克黄金购买一只猪用来过年，但好景不长，频繁交易的黄金由于易磨损、故意囤积、开采难度大等原因，县里的黄金总量根本无法满足人们日常生活交易的需求。

一筹莫展的县长提出一种解决方案：人们将家里的黄金或值钱物品全部到县政府换取纸币，这种纸币上有县长的印章和签名，人们可以通过这种纸币进行交易，一时间，人们欢呼雀跃，纷纷对这位县长的聪明才智称赞有加。就这样经历了十多年，比特县经历了翻天覆地的变化，一跃成为全国十强县。

但就在这个时候，县长却打起了自己的小算盘，他用自己的印章和签名制作了很多属于自己的纸币，一跃成为比特县的首富，更将一部分纸币分发给自己的亲朋好友。久而久之，人们发现，自己手中的钱越来越不值钱，出现

了严重的贬值。经过调查，最终将矛头对向县长，整个县瞬间炸开了锅，人们纷纷要求废除县长的这种特权。

有的人建议成立县政府银行，统一管理纸币的兑换与发放，但人们已经丧失了对政府的信任，也更加排斥这种集中式的存储机构，谁又能保证银行不会乱印纸币呢？就在大家不知所措时，一个名叫中本聪的人发布了一条通告，声称设计了一种去中心化的分布式电子货币系统——比特币，可以解决目前出现的问题。

中本聪提出的比特币系统首先将所有交易记录写入账簿并进行公开，而不再记录每个人的账户余额。交易记录包括付款人、收款人和付款金额，只要通过这些交易记录进行推算，就能获得每个人的账户余额，从而判断交易能否顺利进行。与此同时，为保护个人隐私，每个人不会使用姓名等真实身份，而是使用一串能够唯一标识自己的字符。在这种情况下，每个人都有一个保密印章和印章扫描器，在每笔交易中，付款人和收款人分别用每个人的标识字符表示，并加盖付款人的印章，而这个印记不能肉眼识别，只能通过印章扫描器才能识别。接下来，中本聪在县里招募矿工，这种矿工不是传统的拿着工具去山上采矿的矿工，而是每天在家里花费一定时间完成比特币挖矿工作。同时，根据矿工的挖矿贡献大小，矿工能够获得不同的报酬。

中本聪分别给每个矿工一个新账本，根据目前县中每个人手中纸币的数量，为每个人分配相应数量的比特币，并让各个矿工在各自的账本上记录下这个初始时刻系统为每个人分配的比特币数量，付款人均为比特币系统，收款人分别是每个人的保密印章对应的字符。

接下来，中本聪解释了支付与交易的过程，以老王付给老刘 5 个比特币为例。老王首先向老刘咨询他的标识字符，比如是 0x01a8c3f6，同时老王也有自己的标识字符，比如是 0x2d3c6a89，然后老王写一张单子，内容为“0x2d3c6a89 向 0x01a8c3f6 支付 5 个比特币”，然后老王用自己的保密印章

盖章后交给老刘,老刘在拿到这个单子后,用自己的印章扫描器扫描单子上的印章,发现得到的字符与付款人的字符是一致的,均为 0x2d3c6a89,老刘就可以确定老王签署了这个单子,因为没有任何一个人可以仿造其他人的印章。

在这种情况下,一个关键问题就是老王的账户余额是否足以支付 5 个比特币,而这项工作主要由矿工们完成,即所有的矿工都会收到这个单子,他们将会检查各自手中的账本以确认标识字符 0x2d3c6a89 的余额是否还有 5 个比特币。每个矿工会将收到的交易单记录在各自账本的一页纸上,并将该页纸上内容输入至哈希编码器中,得到一个 256 位的二进制数作为该页纸的唯一标识,如果账本上该页内容被人篡改,则无法对应这个二进制数,这样就保证了内容的安全。同时,矿工也会将账本上前一页的标识以及当前时间值也写在该页上,然后将前一页的标识、该页的标识、时间值和一个随机数(如 56789)输入到一个编号生成器中,若产生的标号前 13 位均为 0,则说明编号有效,并将该编号写到账本的这一页,否则,调整随机数继续尝试有效的编号。

当一个矿工率先得到有效编号时,他会将这页账本的内容告诉所有其他的矿工,其他矿工在收到这页之后,用编号生成器重新计算收到的有效编号,验证正确后将收到的这页账本写入到自己的账本中,以保持所有矿工账本的一致性。所有矿工更新完自己的账本后,都会向老刘说这个交易单是合法有效的,老刘这才放下心来,而第一个挖到矿的矿工也因此得到 50 个比特币(注意,这个数字是示例性的)的报酬。

中本聪在提出不依赖集中处理的比特币概念后,比特县的人民开始慢慢接受并实施,从此,整个县再也没有出现县政府等中央机构投机倒把的现象,而比特币中所使用的机制,正是本章所要介绍的关键技术——区块链。

区块链：让信任成为一种社会共识

我对所有的加密货币及区块链想法及实验充满热情，但我认为，比特币的网络效应，很可能是持久不朽的。

——马克·安德森，硅谷风险投资公司 Andreessen Horowitz 的联合创始人

华山论剑：比特币与区块链

区块链（Blockchain）与比特币（Bitcoin）并不是同一个概念，前者是一项技术，后者是一个系统。如图 8.10 所示，比特币是区块链技术的产物，区块链是依赖比特币的普及才被大家广为接受的，正是因为比特币，区块链在互联网、金融界的重要性才得以慢慢体现。当然，区块链技术目前被数百种的数字货币（如莱特币、点点币等）使用，比特币只是其中之一。可以用 TCP/IP

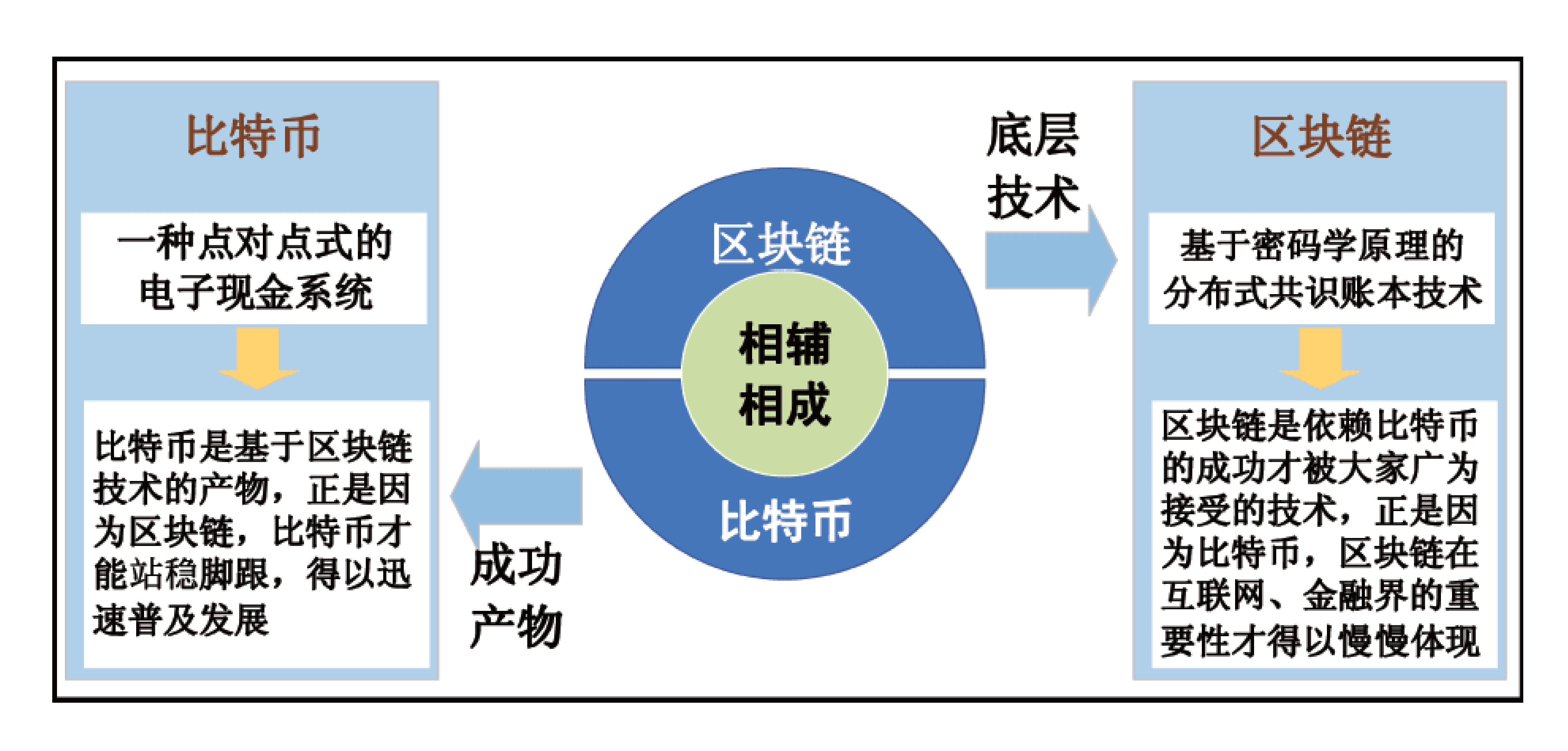


图 8.10　比特币与区块链

协议栈和互联网的关系进行类比，TCP/IP 是一套协议栈，而互联网是采用 TCP/IP 协议栈的实际网络系统。TCP/IP 就好比区块链，互联网就好比比特币系统。

从技术核心来看，区块链是一种基于密码学原理的分布式共识账本技术。区块链利用去中心化的 P2P 技术实现分布式共识机制，完全摆脱了传统的集中处理方式，在保证共识机制的同时，将系统的安全性提升到了一个新的层次。它诞生于第三代 P2P 网络之上，解决了 P2P 网络中的信任和共识问题。

区块链与传统的 P2P 网络一样都具有去中心化的特性。从字面上看，中心化意味着信息都要往一个地方去，显然，如果这个中心出现了问题，整个系统就会受到影响；而去中心化是指系统中的每个个体都是中心，这样只要有节点存在，就像中心化系统中的中心存在一样，整个系统都可以继续运行。那么区块链的去中心化程度如何？从最初的比特币区块链来看，它是完全分布式的，也就是完全去中心化的。

诞生于比特币的区块链技术并不局限于比特币区块链的形式，而是在其基础上不断抽象升华、更新换代。如今的区块链，它的中心化程度可以是弹性的，就如同当初从第二代 P2P 网络到第三代 P2P 网络发展一样，从完全的中心化，到可以根据不同的使用场景允许不同程度的中心化。这种变化使得区块链可以适应纷繁的使用场景，我们可以根据不同的场景，弹性配置区块链，从而在性能、可靠性和安全性之间找到一个最佳的平衡点。

区块链在传统 P2P 网络的基础之上添加了许多新的特性。首先它是不可篡改、不可否认的。建立在传统 P2P 网络之上的信息交换，在信息经过中间节点之后，若中间节点更改了传递的信息，通信双方即使可以通过终端用户检查发现，也无法确定到底是谁从中作梗。在这种网络之下，所传递信息的可靠性需要靠终端用户自己来保证，而 P2P 协议本身可能并不提供这种

功能，这就给用户带来了麻烦。区块链技术将信息的验证融入到了协议之中，不可篡改、不可否认成为了协议本身提供的基本特性，这大大简化了终端用户的操作，同时也提升了整个系统的鲁棒性。

另外一个不得不提的特性是区块链的公开透明性。区块链是完全透明的，所有人而不仅仅是参与比特币区块链构建的矿工和交易者，都可以看到整个比特币从诞生之时起的所有交易信息。当然，这是早期公有链的做法，现在，区块链已不仅仅局限于早期的公有链形式，而是已经衍生出像私有链、联盟链这样的其他种类。这些不同类型的链拥有不同的公开透明性以适应不同的应用场景。

区块链极大的去中心化、不可篡改性、不可否认性、公开透明性共同促成了区块链作为信任基础设施的可行性，解决了参与者之间的共识问题。那么区块链是使用了怎样的技术手段来实现上述这些特性的呢？

把信任留给自己

当前最大、用户最多的区块链就是比特币区块链了，我们通过它来了解一下区块链算法是如何解决信任问题的。

我们知道，整个网络系统中有非常多的用户，他们之间存在非常多的交易，比如张三向李四转账 50 元，李四给王五发 10 块钱红包等等。在传统的交易方式中，资金流转都是经过中介的。你在某电商平台上买东西，钱要先打到某电商的账户上，然后在你确认收货之后，卖家才能从某电商那里得到货款，当然最终解释权归某电商所有。如果某电商侵吞了你这笔钱怎么办？你可以说银行有转账记录，那如果银行与某电商串通呢？你开始头痛了吧？

与传统的交易不同的是，区块链中的每一笔交易都会以广播的形式告知

全网的所有用户，这些交易利用密码学的签名技术都打上了交易双方的签名，抵赖不掉。如果一方不承认两者发生了交易，那么就可以向人民群众求证（发送请求到全网看看大家的记录）。

当网络上的其他用户（节点）收到每一笔交易记录后，首先会对其合法性和有效性进行检查，并将合法的数据记录保存至一个本地的区块中。本地的区块在达到一定条件之后才会被广播到全网，才有可能成为全网共同认可的区块。这种寻找一定条件的过程在比特币的系统里，称之为挖矿。

随着用户间交易的不断产生，本地区块中可能包含了多个交易记录，在某一刻用户可能找到了那个需要达成的条件，那时候用户就可以将他本地的区块广播到全网，这些区块可以包含全网用户在一定时间段内的交易。比特币区块链的挖矿过程是十分消耗 CPU 资源的，需要不断地进行大量计算，这也是人们从密码学理论上认为其难以被修改的原因，是构成比特币区块链信任的基础。在区块被广播到全网之后，大家都对该区块进行检查。由于标准的检查过程是相同的，就像国家标准一样，达到了就是达到了，没达到就是没达到，只要区块是有效的，那么它就会被全网的用户都接受，这就是共识过程。

聪明的读者可能发现了另外一个问题，用户为什么愿意做这种为人民服务的事情呢？原因很简单，一来用户记录这些全网的交易是要收税的（交易费用），二来既然我们称寻找达成某个条件的过程为挖矿，自然就可以将挖到的“矿”作为自己的收入，这个就是比特币。全网所有矿工在收到有效的区块之后，都会停止在当前区块的挖矿工作，因为当前的矿被别人挖走了。在将这个区块追加到上一个区块的尾部之后，矿工们又在最新区块的基础上开始了它们新的挖矿之旅。就这样一块接着一块就形成了区块链。

图 8.11 描述了区块链的结构，每个区块中含有多个数据记录，这些数据

记录按照一定的组织形式存在于区块之中，而这些数据记录的哈希值构成能够唯一标识区块的 ID。每个区块会记录上一个区块的 ID，同时又将区块生成的时间作为时间戳嵌入到区块头部，这样，区块可以按照顺序链接到一起，从而形成区块链。

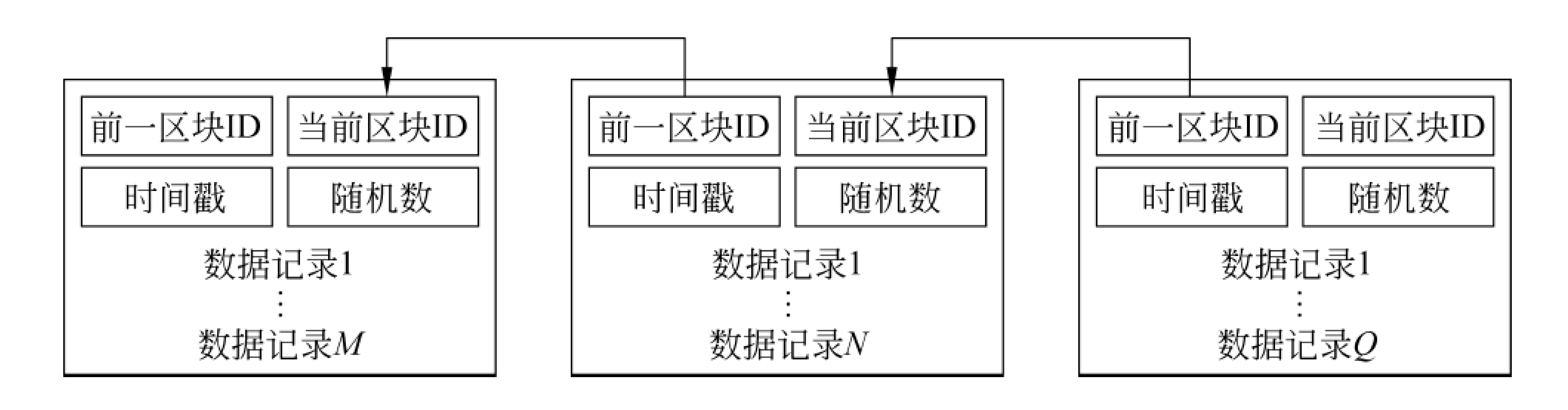


图 8.11　区块链结构

由于各个节点是分布式并行工作的，因此区块链在实际运行过程中有可能出现**分叉现象**（如图 8.12 所示），即在同一个区块的基础上产生多个后续区块，这一方面会影响区块链的稳定性，另一方面则会浪费计算资源。

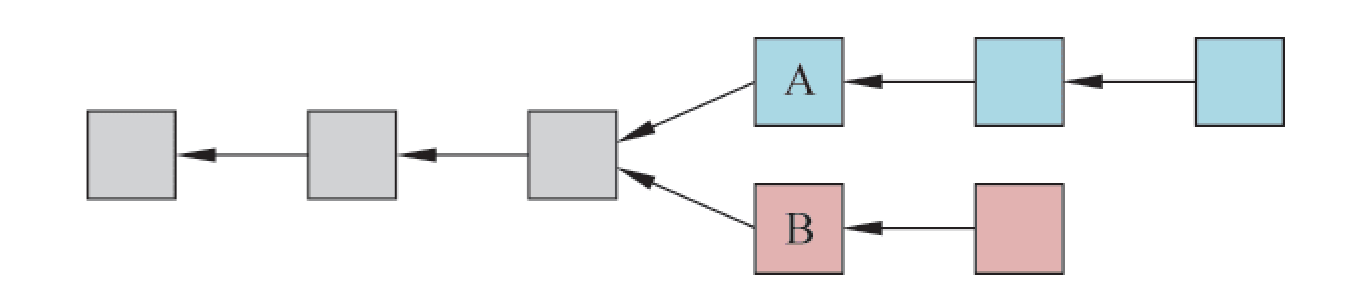


图 8.12　区块链的分叉现象

在区块链中，共识机制可以很好地解决分叉现象，即长度最长的分支将成为主分支，并纳入到区块链的主链中。这样，在区块产生及区块链延长过程中的小分支将会淹没在历史的长河中，这有利于保证区块链数据的一致性和安全性。

区块中保存的数据记录以默克尔树的形式组织在一起，如图 8.13 所示，所有的数据记录存在于树的最低端，经过哈希计算，每条数据记录对应于一个哈希值，相邻哈希值两两结合形成新的哈希值，就这样由底向上，当前区块

ID就是所有交易记录的最终哈希值。数据记录的这种组织结构有利于防止数据信息篡改,如果其中一条数据记录遭到恶意修改,对应的哈希值也会发生变化,并最终使得当前区块ID发生更改,而被更改的区块在其他用户那里是不能验证通过的。

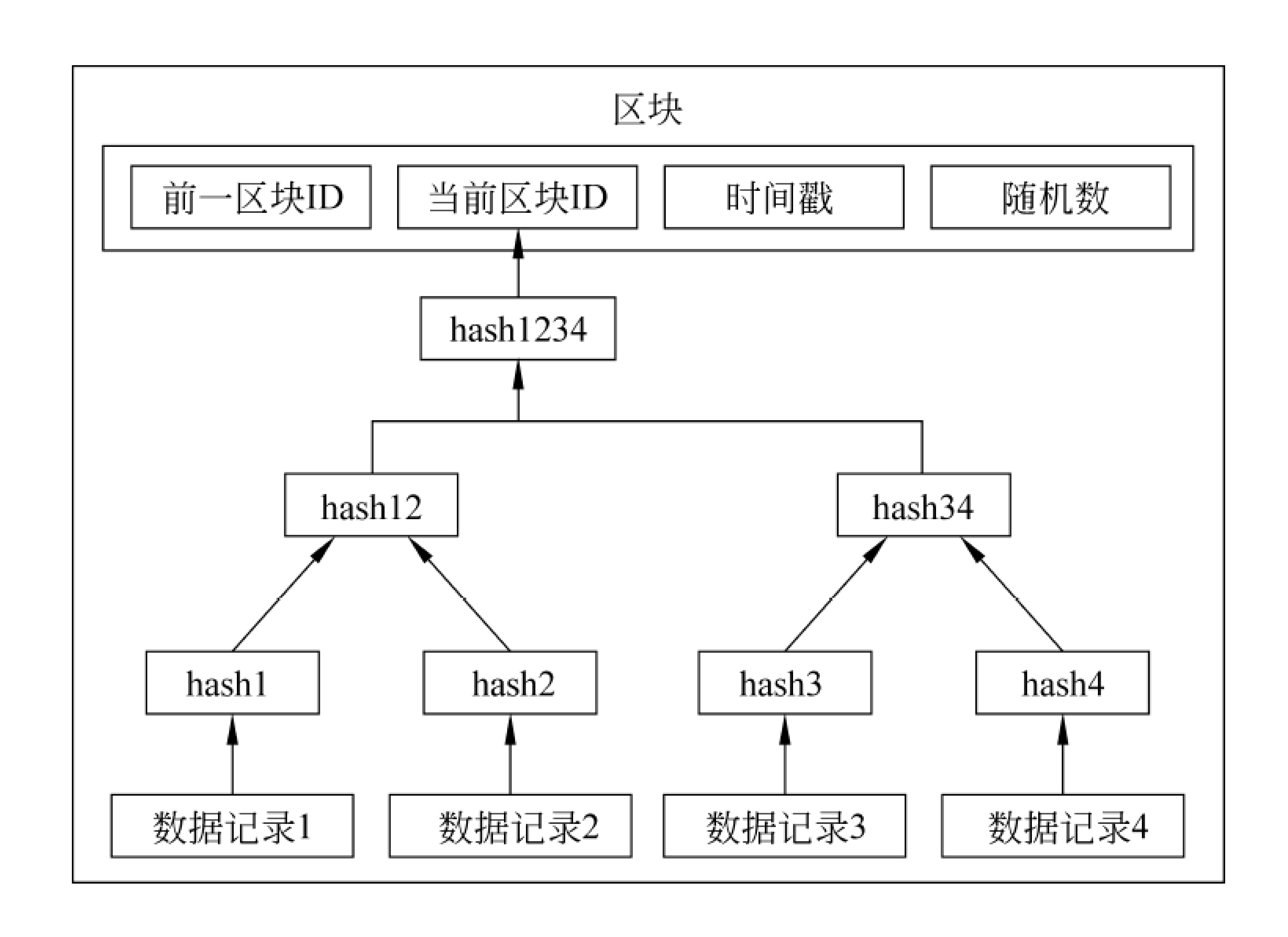


图8.13　区块的结构

区块中的数据记录如图8.14所示。在比特币中,数据记录的表现形式为交易(Transaction),它利用密码学中的非对称加密技术来保证其安全。在非对称密码学中有两个密钥,即公钥和私钥。用户将公钥发布给其他人,而将私钥秘密保存起来。公钥加密技术提供两个基本功能:加密和签名。经公钥加密之后的信息只有私钥可以解密,由于私钥只有密钥产生者知道,所以其他人不能破解被加密的信息。而私钥还可以用来签名,其他任何人都可以用该私钥对应的公钥来验证该签名,由于私钥的私密性及唯一性,该数字签名就像人们常用的按手印和手动签名一样具有不可抵赖性。在区块链中,每个节点都有一对公钥和私钥。以比特币为例,区块链中的交易过程如图8.14所示,该过程包含交易签名和交易验证两部分。

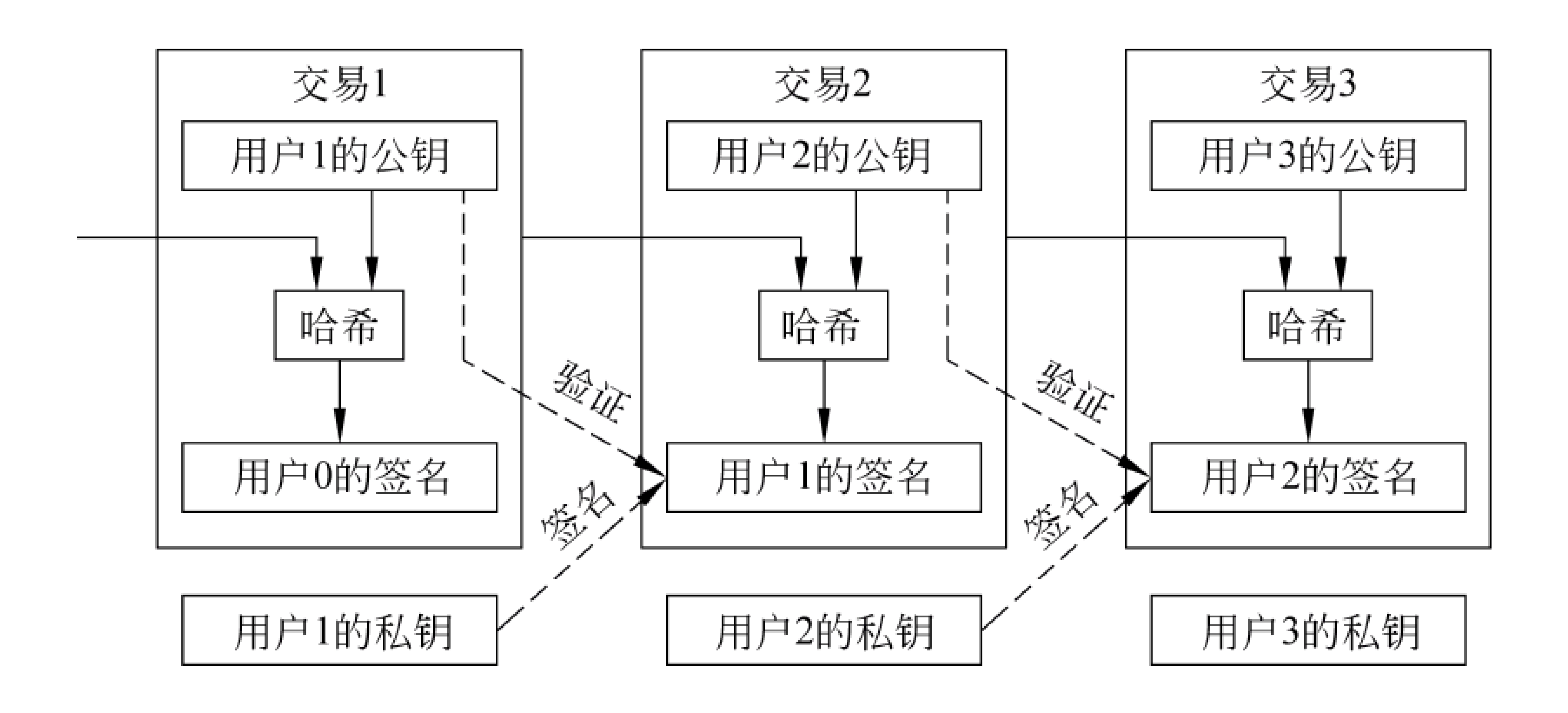


图 8.14　区块链交易示意图

假如用户 1 与用户 2 存在某种交易（即图中的交易 2），用户 1 首先进行交易签名：利用用户 2 的公钥对上一个交易（即交易 1）进行哈希加密，得到哈希值；然后用自身的私钥和该哈希值对该交易（交易 2）进行签名，并将该签名嵌入到该交易中，最后将该交易向全网广播。全网节点在收到该交易后，首先进行交易验证：利用用户 2 的公钥对上一交易（交易 1）进行哈希加密计算得到哈希值 X；利用用户 1 的公钥对该交易（交易 2）中的签名进行解密，得到解密值 Y；如果 $X=Y$，则证明当前交易的发送者和接收者确实分别是用户 1 和用户 2，即用户 1 和用户 2 确实要进行本次交易。经过交易签名和验证后，当前交易才能加入到区块中。

相比传统的解决方案，区块链技术在不依赖第三方介入的情况下，成功地解决了双花问题（double spending，又叫双重消费）。所谓的双花问题，是指一个用户在账户余额有限的情况下，同时向两个或以上的人支付总共超过账户余额的金额。依赖第三方仲裁该用户的余额是否足以支付是传统的解决方案，而区块链中通过广播每一笔交易以及在交易中添加时间戳更好地解决了双花问题。

聚焦区块链的共识

在区块链技术中,共识机制是实现用户间信任的技术核心,也是保证全网节点就区块信息达成一致共识的关键技术。在共识机制下,区块链的延长具有一致性,区块信息具有安全可靠性。如图 8.15 所示,当前在区块链系统中,交易记录和新区块都会广播至全网,使之成为全网节点的共同知识,这是区块链共识机制的前提,而区块链的共识机制主要是建立在工作量证明、权益证明和股份授权证明的基础之上的。

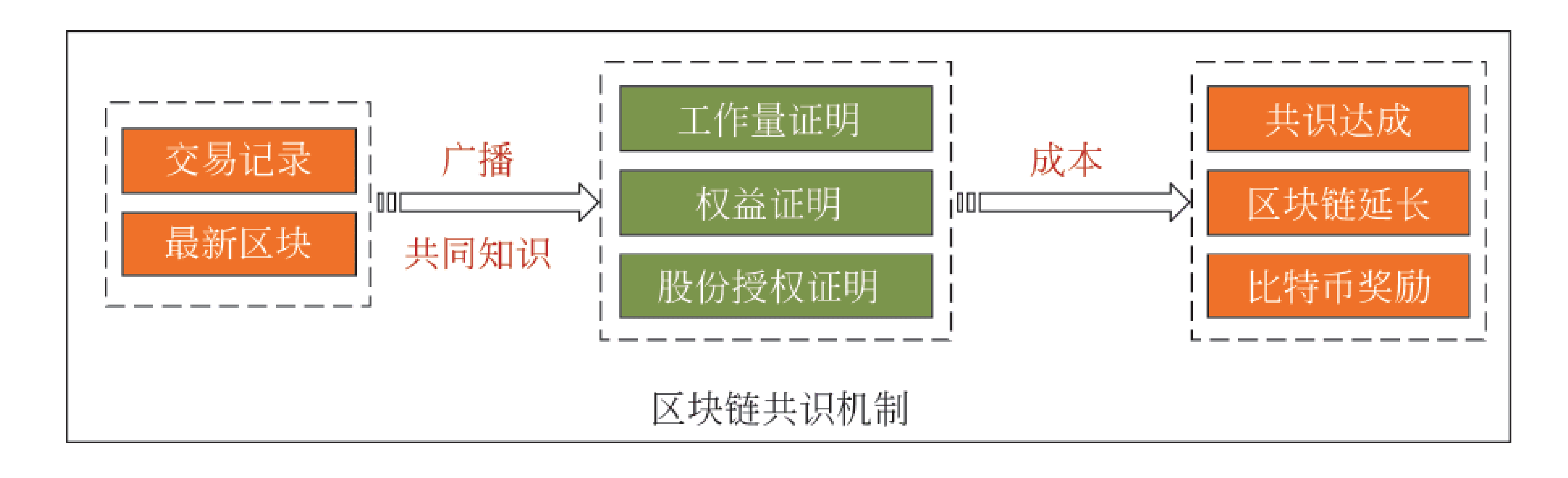


图 8.15　区块链共识机制

工作量证明

工作量证明(proof of work,POW)是建立在以下两个问题的考量上:一是,如果随便一个节点在短时间内都能将新区块加入到区块链中并广播给全网,那区块链岂不是存在太多的分叉,以至于短时间内根本无法确定主分支?二是,全网节点为什么要拼命执行这套机制?无利不起早,如果没有奖励机制,谁会愿意干这事呢?在区块链中,工作量证明就回答了这两个问题。

首先,工作量证明使得每个节点需要付出一定的努力和代价(如 CPU、硬盘和电量等)才能形成新区块,并链接至区块链中。这体现了节点的工作

量，工作量越大，越有可能形成新区块。在区块链中，每个节点通过调整区块中的随机数，并利用哈希计算得到符合条件的工作量证明数值。这个过程每个节点需要不断尝试不同的随机数，而寻找合适的随机数完全是一个概率事件，所以极短时间内根本无法完成，也恰恰由于这个原因，区块链的分支现象会在一定程度上减少。其次，在区块链中，工作量证明也是一种激励机制，它会对生成新区块的节点（即矿工）给予一定奖励，在这一定程度上提高了矿工的积极性。

讲到这里，让我们再重新回顾一下拜占庭将军问题，如何在一个互不信任的分布式群体中达成信息共识？

区块链技术通过一个看似简单的办法解决了这个问题，它为发送信息加入了成本，目的是为了降低信息传递的速率，并加入了一个随机元素以保证在一个时间里只有一个将军可以进行广播。

这里加入的成本就是“工作量证明”，并且它是基于计算一个随机哈希算法的。以比特币为例，哈希算法要得到一串 64 位的随机数字和字母的字符串，就像这样：

d70298566aa2f1a66d892dc31fedce6147b5bf509e28d29627078d9a01a8f86b

尽管单个哈希值用现在的计算机可以几乎立刻计算出来，但只有前 13 个字符是 0 的哈希值结果可以被比特币系统接受成为“工作量证明”。这样一个 13 个 0 的哈希值是极其罕见的，大致需要花费整个比特币网络 10 分钟的时间来找到一个。在一台网络中的机器随机地找到一个有效哈希值之前，上十亿个的无效值会被计算出来，这就是减慢信息传递速率并使得整个系统可用的“工作量证明”。下面是一个例子：

f51d0199c4a6d9f6da230b579d850698dff6f695b47d868cc1165c0ce74df5e1

d70298566aa2f1a66d892dc31fedce6147b5bf509e28d29627078d9a01a8f86b

119c506ceaa18a973a5dbcfbf23253bc970114edd1063bd1288fbba468dcb7f8

在找到一个有效值之前，成百万甚至上亿个更多的类似上面这样的字符串被计算出来，直到下面这样的有效值被发现：

00000000000084b6550604bf21ad8a955b945a0f78c3408c5002af3cdcc14f5

那台发现新的有效哈希值的机器（或者说是拜占庭问题中的某位将军），把所有之前的信息放到一起，附上它自己的信息以及它的签名，然后向网络中的其他机器广播。只要其他机器接收到并验证通过了这个13个0的哈希值和附着在上面的信息，它们就会停止它们当下的计算，使用新的信息更新它们的区块链，然后把新更新的区块链作为哈希算法的输入，再次开始计算哈希值。哈希计算竞赛从一个新的开始点重新开始。如此这般，网络持续同步着，所有网络上的计算机都使用着同一版本的总账。

分析表明，POW共识算法其实是一种**概率性**的拜占庭协议，在不诚实节点总算力小于50%的情况下，当任意两个诚实节点的本地链条截取K个节点，两条剩下的链条的头区块不相同的概率随着K的增加指数递减；当不诚实算力非常小时，才能使大多数区块由诚实节点提供协议的正确性。虽然工作量证明在区块链技术中意义重大，但它是以计算机资源和能量的消耗为代价的，这显然是一种浪费。同时，如果某个节点或多个联合节点拥有较强的算力，它（们）甚至可以左右整个区块链，这也会为区块链的安全埋下隐患。

权益证明

由工作量证明驱动的区块链技术是以资源与能源的巨大浪费为代价，而权益证明（proof of stake，POS）解决了这个问题，同时又不会影响区块链中的激励机制。权益证明的基本思想是用节点的权益值代替工作量来获取记账权利并获取激励收益。权益证明在有限的范围内尝试随机数，能够避免资源与能源的消耗。权益体现了节点对特定数量数字货币的所有权，用币天（coin days）表示，表示币数与最后一次交易的时间长度的乘积。在解决分叉

问题上，区块链选择消耗最高币天的分支作为主分支，这样，任何一个试图改变分支的一个或多个节点都会付出更多的币天为代价，这样的难度显然是更大的。在现有的区块链中，有一些共识机制是采用工作量证明和权益证明结合的方式，这样在达到节约资源与能源的基础上，又把强算力节点联手影响区块链安全的问题解决了。

股份授权证明

股份授权证明（delegated proof of stake，DPOS）是权益证明的演化机制，类似于股份公司中的董事长选举，基本思想是节点拥有的权益越多就有越多的投票权，通过投票选出得票最多的前 100 个代表节点，每个节点轮流负责生成区块，并将一定比例的激励平均分发给所有代表节点。作为代表节点，必须保证实时在线，为大家提供良好的区块生成服务，否则在下次投票选举中很有可能丧失代表资格。这种方式实现节点轮流记账，大幅减少了参与验证与记账的节点数量，在实现快速共识验证的同时又不影响区块链数据的安全性。

区块链技术：带动赛博经济进入智能经济时代

近几年，包括美国、英国、中国、俄罗斯等各个国家纷纷对区块链技术作出响应，对本国区块链的未来作出规划。

2014 年，美国加州签署法案，将使用比特币交易确认为合法行为；2014 年底，美联储发布了一份改善支付系统的白皮书：“比特币技术目前并不足够成熟，但是我们有兴趣进一步探索该市场”；2015 年 6 月，纽约州金融服务局公布了简称为“比特币牌照”（BitLicense）的《虚拟货币监管法案》，它是美

国第一个专门为虚拟货币量身打造的监管规则。

2016年1月,英国央行行长Mark Carney发表题为《分布式账本技术:超越区块链》的报告,体现了英国对区块链技术的重视;同年6月,他在演讲中展望了区块链和互联网金融变革全球金融系统和英国经济的方式;7月20日发布最新报告指出,用(英国)央行发行的数字货币替代30%流通中的货币,可以“把GDP提升三个点,并永久保持”,同时“大幅改善央行稳定商业周期的能力”。

2016年10月,俄罗斯央行开始测试基于以太坊的区块链原型——Masterchain;Tinkoff银行行长兼首席信息官Viacheslav Tsyganov说:“我们与合伙人一起开发的基础设施已经经历了一次测试,证明了区块链技术的最大潜力,及在搭建银行间通信渠道的适用性”。

2016年1月,国际货币基金组织发表《虚拟货币与超越:初步探讨》,阐述了虚拟货币在包括支付和价值转移,特别是跨境支付和价值转移方面有着非常大的潜力,认为能够在推动普惠金融发展方面发挥作用,其背后的技术所引发的变革将远远超过虚拟货币本身。

2016年1月,中国人民银行对区块链等数字货币技术显示出了高度肯定,首次将发行数字货币作为央行的战略目标;12月27日,国务院将“区块链”写入“十三五”规划;2017年3月,全国人大代表、中国人民银行营业管理部主任周学东提出建议:(1)对境内比特币交易平台应当包容、暂不取缔;(2)在短期内必须明确比特币交易平台监管红线,严格监管;(3)建立负面清单,做好风险防范和化解工作。

在赛博新经济时代,区块链是信任的基础设施,是信息互联网到价值互联网的跳板。价值互联网是指人们在互联网上可以像传递信息一样方便快捷地传递价值,例如资金。在价值交换中,安全是一个最值得关注的问题,而区块链恰恰可以让两个互不信任的人之间进行安全可靠的交易。作为信任

基础设施，区块链技术将令未来互联网上的信息与价值交换，不再必须经过第三方的介入。区块链就像是高速公路，大家都可以使用，而且大家可以信任它，不必担心它的安全问题。同时在无中介的情况下，点到点的价值交换将节省大量的手续费，更加方便快捷。区块链可以被应用到金融、物联网、商业、安全、信誉系统等各个领域，其未来是不可估量的。

随着人工智能的发展，智能家庭、智慧城市、智能交通等新的产品及商业形态也在不断涌现，区块链技术也加速了这样一个万物互联时代的到来。在区块链基础上，人们可以利用智能合约构建各种各样的智能系统，满足不同需求。智能合约其实早在 20 世纪 90 年代初期就被提出来，但是一直苦于没有可信的执行环境，天生带有信任特性的区块链出现填补了这一缺口，使得智能合约得以被应用。赛博新经济时代也将在利用智能合约的区块链的助力之下，迈向赛博智能经济时代。在赛博智能经济时代，在完全没有人工干预的情况下的物物交易将成为一个非常普遍的现象，人们将从烦琐的小事中解脱出来，例如，洗衣液使用完后，洗衣机可以自动购买，电冰箱可以自动购买被日常消耗掉的水果蔬菜等。

当然，区块链想要发展，仍然需要解决一些问题。在 2016 年 Money 20/20 会议上，以太坊创始人 Vitalik Buterin 就曾指出："那些声称区块链是绝对不可更改的论断实在太可笑了。因为事实就是，当时只要你有 4000 万美元，你就能对以太坊发动 51%攻击"。目前，区块链主要在可扩展性和安全性以及隐私泄露、共识算法的性能等方面依然存在挑战。

从交易频次来看，比特币区块链每秒最多处理交易次数的极限为 7 笔。相比之下，Visa 的平均速度为每秒 5000 次交易，峰值大约每秒 60 000 次交易，支付宝"双十一"的峰值为每秒 120 000 次交易。不仅如此，区块链的分布式特征导致其流量必须广播至全网，这将带来巨大的性能开销和资源浪费。

同时，区块链技术在安全性方面也并不是无懈可击，攻击者的计算能力超过全网的51%就可能操纵整个区块链，从而达到篡改交易记录、截断区块链正常延伸的目的。虽然在庞大的网络系统中，拥有51%的计算能力或控制51%的节点并不现实；有分析显示，这种方式的攻击即使会成功，也要花费大约5000万美元的成本，这种攻击防御的不对称性破坏了攻击者的意图，在一定程度上保证了区块链的安全。

此外，在区块链底层技术层面，非对称加密技术也并不是那么牢不可破：2017年2月23日，Google在其安全博客上公布找到了世界首例SHA1碰撞，标志着SHA1不再安全；而区块链用到的SHA256技术会不会在不远的将来被破解？一旦SHA256被攻破，比特币区块链便不再安全。再如，区块链底层采用的椭圆曲线加密会不会在快速发展的量子计算机面前"漏洞百出"？这些都将成为区块链的安全隐患。

未来，当智能合约与区块链结合，一方面意味着方便快捷，另一方面也使各种各样漏洞的利用变得更加容易。以2017年5月爆发的比特币勒索病毒为例，攻击者利用Windows系统的漏洞来入侵用户系统，对用户的文件进行加密，从而以此威胁用户支付比特币到特定账户（即利用公钥生成的地址）。用户支付比特币之后，受害者系统中的病毒通过自动扫描区块链来检查用户是否支付，以决定是否恢复被加密的文件。用户感染病毒后的加密、勒索、检测的过程都是自动执行的，不需要攻击者的参与，这很好地隐藏了攻击者。即使后期攻击者需要交易指定账户的比特币，在其有所防备的情况下，攻击者可以通过许多方法，比如比特币的混合服务使得对特定账户的追踪变得极其困难，这就导致了攻击者可以肆无忌惮进行勒索。虽然这并不是比特币区块链本身的安全问题，但从一个侧面反映了区块链的出现使得以前不可能的攻击方式成为现实。就像新事物总会带来新问题一样，如何处理好区块链带来的连锁反应，也势必伴随其发展过程的始终。

第9章　未来：赛博智能经济

算法成为未来经济系统演变的重要推动力量，人类正在进入一切皆可计算的时代。在算法强大的力量下，赛博世界中产生的海量数据和信息不再是一团乱麻，算法可以轻易找到隐藏在“乱麻”中的“线头”，让这些数据和信息变得有秩序，而这正是经济增长的基础。算法还是人们构筑可信任机制的基础，是经济运行安全稳定的保证，不仅现在，还包括未来。算法不仅仅是社会经济运行的添加剂，更是凌驾其上，定义和管控着整个经济运行过程，形成了当下的赛博新经济，并推动赛博新经济朝着赛博智能经济的方向发展。如果说算法重新定义了今天这个世界的经济秩序，那么未来它也必然会开启一个赛博智能经济的黄金时代。

赛博智能经济的雏形

我们需要万分警惕人工智能，它们比核武器更加危险！

——埃隆·马斯克，企业家

计算机科学领域的专家预测2049年人类的工作生活是这样的：

人人都是程序员

“把按钮A填上红色，放在左上角，上面写‘注册’……”黎桐对着手机说道，“注册用户数量暂定五千万，软件名称叫XX……”说完黎桐又对着手机屏幕做了几个手势。如果这段对话发生在今天，那么最可能的场景便是产品经理在和程序员在沟通。但这段对话发生在2049年，黎桐是一名程序员，准确地说，是一名兼职程序员。她并不是在跟谁说话，而是在一个人编写程序。是的，她用人类语言描述出了她想做的程序，完成了“源代码”的编写。然后又用几个手势搞定了建立数据库、租赁服务器和域名等操作。这对于今天的我们来说也许很难理解，一个不懂任何算法、数据库和服务器知识的人，是如何能通过几句话、几个动作就完成了一套可执行的程序？

在2049年，设计师在写程序，医生在写程序，会计在写程序，作曲家也在写程序……写程序不再是外行人望尘莫及的复杂工作，因为到了那时，人人都是程序员。程序的编写形式也将变得多样性：人们可以写程序，说程序，画程序……任何一个有需求的人都可以直接与计算机沟通，创造自己想要的程序，让计算机为自己做事。每个人都可以像用计算机写一段文档一样简单地让它为自己创造想要的产品和服务。

智能的人联网

从你睡醒睁开双眼的那一刻，你已经生活在一个智能机器人充斥的环境中：你的家本身就是一个智能机器人，智能卫浴会为你自动调整洗浴水温，智能厨房会为你自动烹饪早餐，出门上班时，交通工具会是无人驾驶的机器人汽车，当你走进办公室，你的智能桌子会立刻感应到，然后为你打开邮箱和一天的工作日程表。在一个人物联网的时代里，你的手表、项链、戒指、眼镜，洗衣机、冰箱等一切都是智能化，它们不需要人的介入就能准确运行，算法真正地实现了物的自主“思维”，也就是说，新算法的产生和实现不需要人类的干预，由控制算法根据设定的规则自主决策。

相比未来的赛博智能经济，目前的赛博新经济可以说只是这一切的雏形。之所以如此，是因为现在算法仍然是由人类设计和实现的。人类由于自身知识结构和认知能力的局限，在理解和认识我们的物质世界时是不全面的，还存在不少模糊不清楚的地带。况且，人类一直是在用世界上已经存在的东西来帮助我们理解这个世界，这仿佛是一个悖论。人类的数学家、计算机科学家在设计算法时，往往需要模型的帮助。模型是人类对真实事物或现象的形式化描述。不幸的是，在这样的描述过程中，由于人类的大脑能够处理的信息量也是有限的，最终构建的模型往往是对真实世界的简化。例如，最简单的线性一元一次方程对大部分人来说不是什么难事，但随着未知量数量的增加，并且变量之间的关系是非线性的时候，人们往往会束手无策。《增长的本质》一书中曾提出过一个叫做“人比”(personbyte)的度量单位，这个单位的含义是，“一个人的神经系统所能接收的最大信息量”。基于这个概念，一个人只能积累一个人比的信息，如果他/她试图获取的信息超过一个人比，则需要其他人的协助。人比这个概念说明了人们积累知识的容量和能力是有限的。再回到上面模型的例子，人们在理解事物或现象时，如果所需的信息小于一个人比，就可以充分认识该事物或现象。不过，我们身处的世界

又是复杂的,认识其中的事物或现象所需的信息往往会超过人比的限制,这时,人们往往不得已采用简化处理的办法,而这正是人类对世界的认识和理解出现偏差的根源。

人们的认知能力受本身生物属性的限制,在有限能力的情况下设计实现的算法,不一定是错误的,但一定存在改进和优化的空间。此外,目前的算法存在一个最大的软肋——缺乏自主性。当前的算法是人类设计和实现的,算法的处理流程和所谓的"智能"其实都是人类思维的衍生,从本质上来说,并没有摆脱人类的局限性。可以说,现阶段的初级算法,只能应对社会经济这个巨型复杂系统中很小的一部分。即使这样,现代经济依然在算法的推动下进入了赛博新经济时代,同时展现出了巨大潜力,也为未来的智能经济时代带来了无限的想象空间。

从微软公司的小娜、小冰和苹果公司的 Siri,到 2016 年战胜韩国棋手李世石的围棋程序 AlphaGo,都展露出未来智能算法强大能力的一角。2017 年 5 月底,围棋人机大战第二季在浙江乌镇开始。大战之前,人们都很期待,这次 AlphaGo 会下出怎样的围棋?Google 会不会派出号称"从不接触人类的棋谱,完全靠自主训练成长"的版本?这样一个"纯净的完全没受过人类污染的"AlphaGo,能否完全颠覆人类的围棋理论和认知?最终结果很具冲击力,AlphaGo 以 3∶0 的战绩零封人类顶尖高手柯洁,其中很多手棋都让众高手叹为观止。

据说,去年战胜李世石的是 AlphaGo 的 v18 版本;另一个名为 Master 的 v25 版本已经在网上连胜了人类 60 局,其学习成长的速度再次引爆人类的眼球。在这些对局中,Master 已经下出很多让人类棋手看不懂的招式,而且越来越多的人开始模仿和学习 Master 的某些套路。还以柯洁为例,虽然不敌 AlphaGo,但经此一役,柯洁脱胎换骨,截至 2017 年 7 月 17 日,柯洁已经对人类选手取得了 22 连胜,彻底坐稳了一"狗"之下,万人之上的宝座。

下面来看看 AlphaGo 曾经用过的几个“奇招”①，从中可以体会算法的智能。

招式一：

图 9.1 展示的是李世石与 AlphaGo 第二盘对局的初始阶段。这里，AlphaGo 执黑在棋盘右下角的托退定式还没有完成就开始脱先，到上边拆边构成了“中国流”。李世石执白在左边星位夹击后，黑棋走了图中▲所示的尖。在传统理论中，这手棋可以说是绝对先手，很多职业棋手都会选择稳一稳，到最好的时机再下这一手，以免早早浪费一个劫材。“保留”也是区分棋力高低的一个传统方式。然而，自从 AlphaGo 走出这招以后，人们很快就认可了，觉得这是简单明快定型，以免将来走不到这个先手。

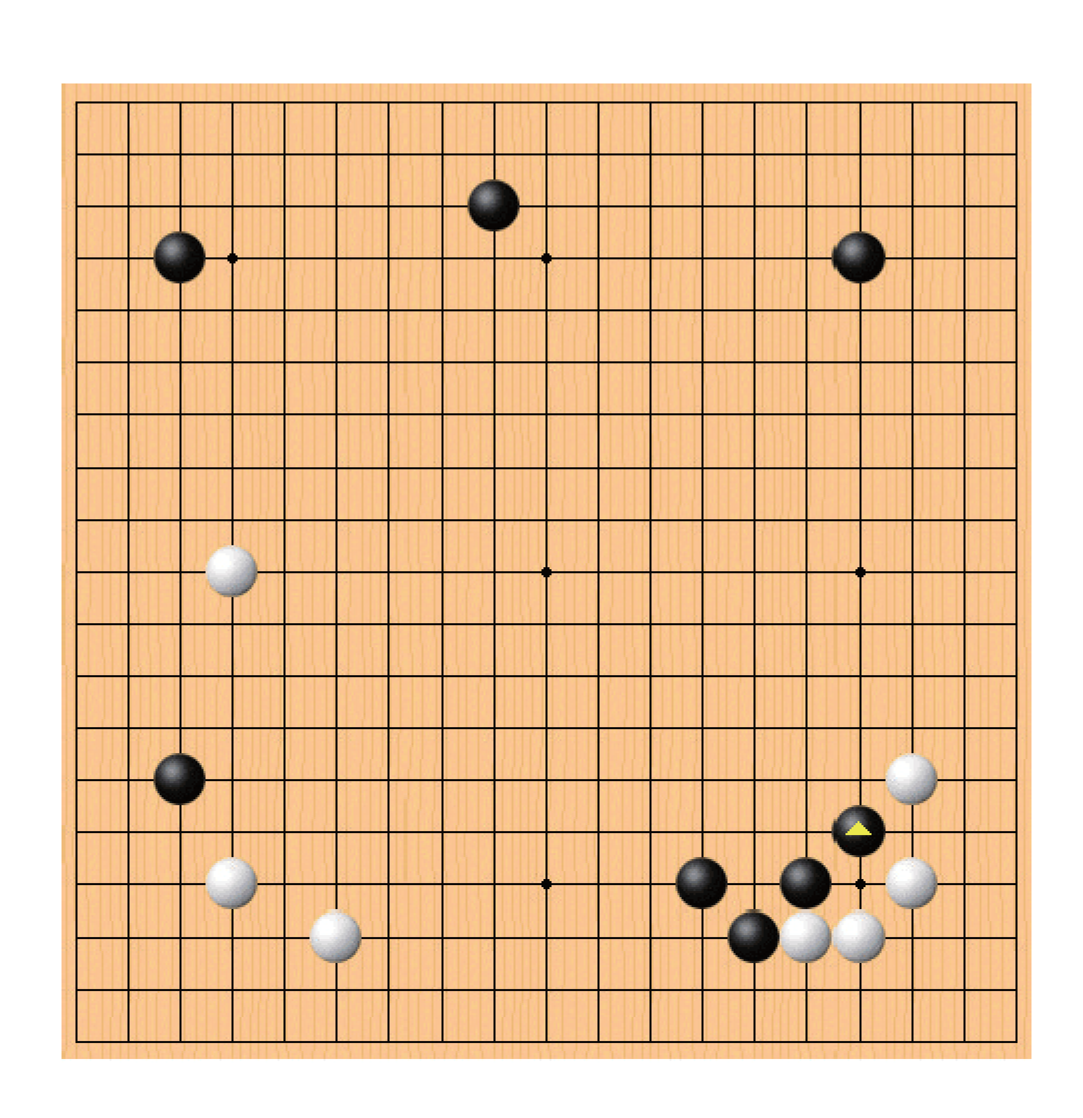


图 9.1　简洁明快的“尖”

招式二：

图 9.2 展示的是第一次人机大战时，AlphaGo 最令人难忘的一手，即图中的▲。这一步出乎所有人意料，完全不属于人类围棋思维，极具视觉冲击

① 腾讯体育：http://sports.qq.com/a/20170426/007550.htm。

力。在右边完成几手交换后，AlphaGo粘住左边的两颗子，然后在人类认为绝无可能占到便宜的地方主动开战，最后大获成功。

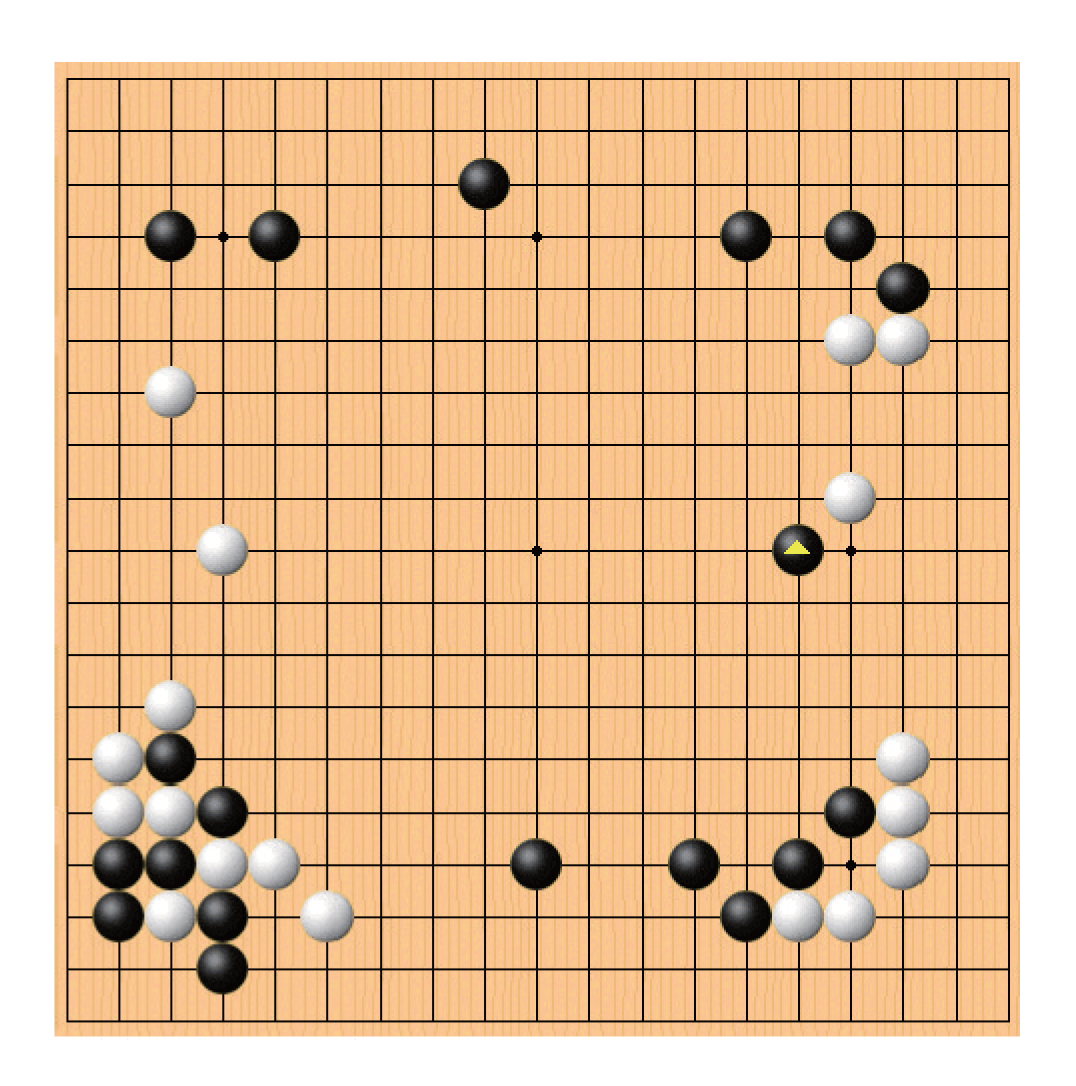


图9.2　极富冲击力的一手

招式三：

图9.3展示的是Master执黑对阵中国棋手檀啸的序盘。其中黑11的“刺”，只在围棋大师吴清源的棋谱中出现过。赵治勋曾评价这手棋是“简明

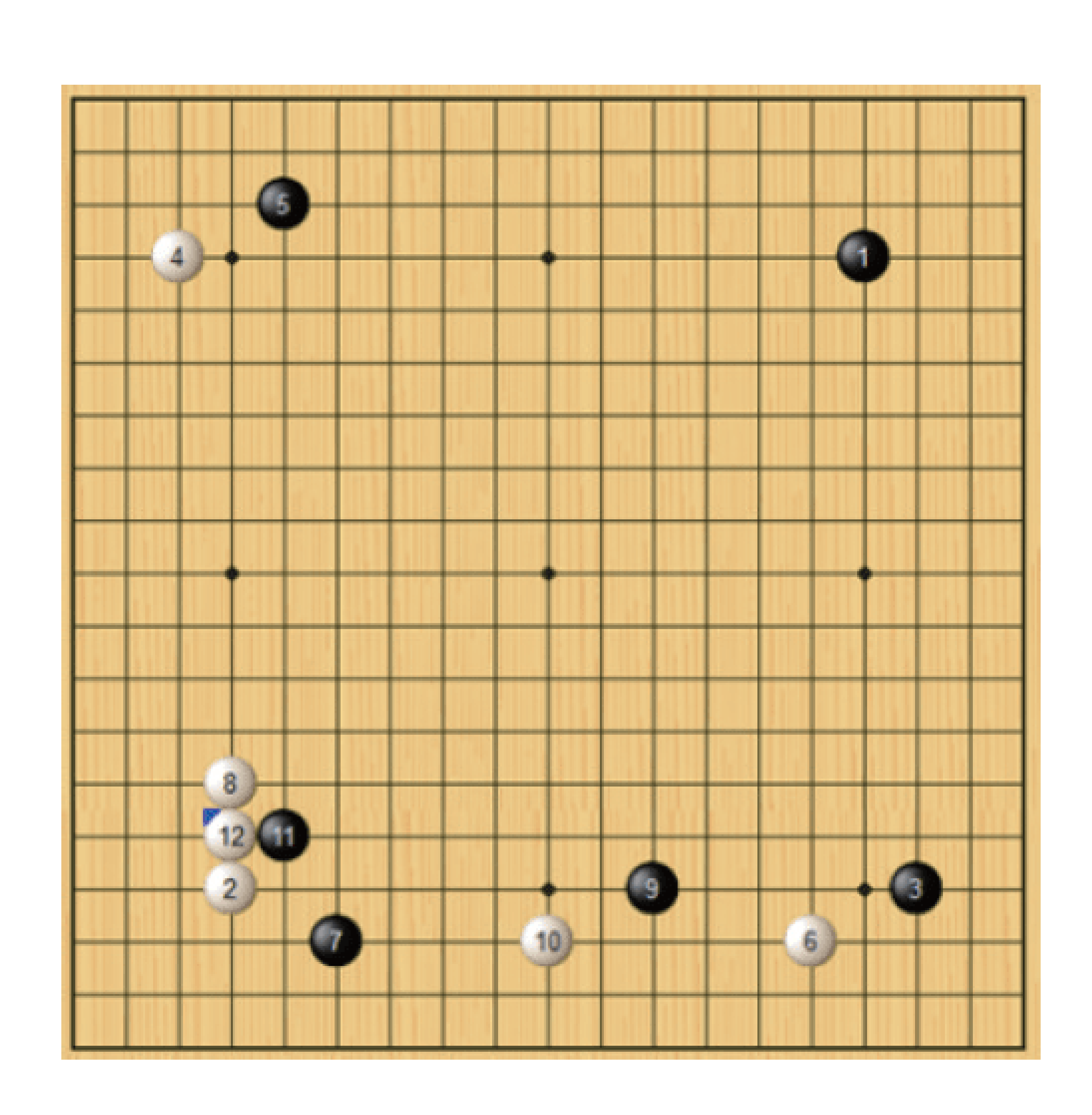


图9.3　与吴清源心灵相通的一手

快速的棋风，喜欢不容分说的转换”。在吴大师之后，还没有棋手敢走这一手，直到 Master 出现。难道说冥冥中 Master 与大师心灵相通？

从 AlphaGo 在这几次对局中的表现看，人工智能在飞速发展，其进化的速度远远超过了人们的估计，在智能化的道路上不断前行。

人类无法理解算法带来的新知识

不是我创造的，我就不能理解。

——理查德·费曼，物理学家

2008 年，《连线》杂志前总编辑、无人机公司 3D Robotics 首席执行官克里斯·安德森（Chris Anderson）写下了这样一段话：“海量数据的可用性，以及用来分析这些数据的统计工具，提供了一种认识世界的新方式。相关性取代因果关系，即便没有一致的模型、统一的理论或者任何机械论解释，科学也能够前进。”这段话当时在学术界引发了广泛而激烈的争论。有学者在某分子生物学期刊上发表文章回应：“……如果我们停止去寻找模型和假说，那我们所做的还是科学吗？答案显然应该是‘不’。”

到如今，这段争论已经过去 9 年，当年的争议在今天看来结论已经显而易见。在全球网络化，以及强大的计算机硬件的助力下，计算机算法取得了飞速的发展，已经使得计算机能够不用模型就能运作。不仅如此，计算机算法还能够自己生成模型，人们需要做的仅仅是给它提供数据，尽管那些模型看起来难以理解，不太像人类构造出来的。但在当下这个“人工智能和机器学习”重新掀起新浪潮的时代，这样的情况正在变得越来越普遍。

我们举算法构造模型的例子。基因表达式编程 GEP（Gene Expression

Programming)是借鉴生物遗传基因的结构和表达规律提出的一种自适应演化算法。GEP 中用字符串表示染色体,这些字符串代表数学表达式,可以解码为表达树。比如,染色体“sqrt * + * a * sqrtabc/1-cd”就代表图 9.4 中展示的数学表达式及其表达树①。

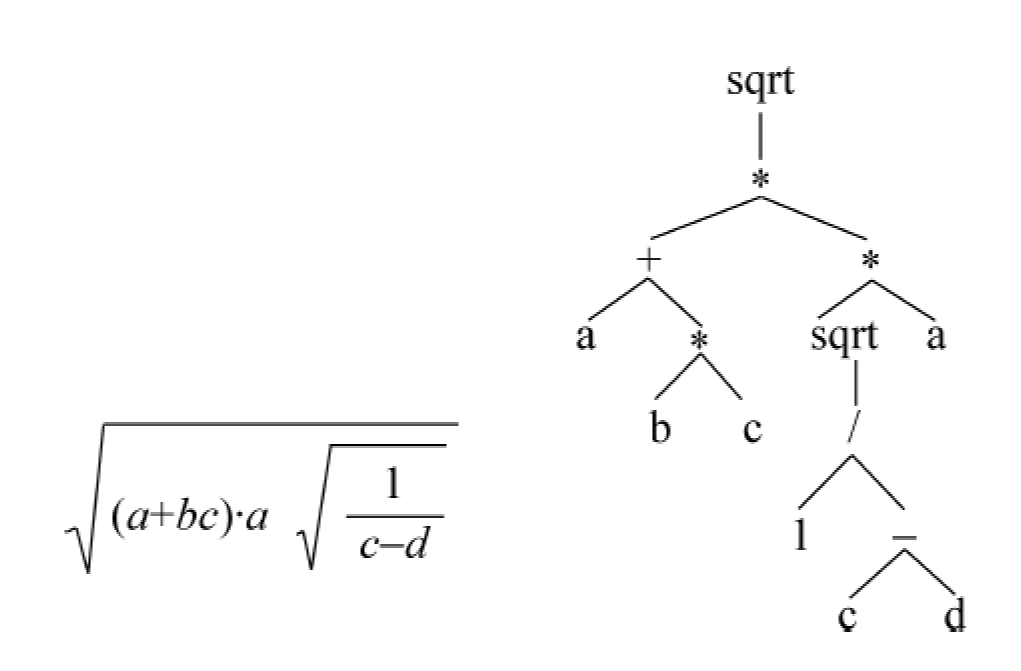


图 9.4　染色体代表的数学式及其表达树

GEP 的计算过程就是这些代表“染色体”的字符串仿照自然界中生物进化过程,采用选择、变异、交叉等操作算子,使染色体开始一代代向前进化,也就是其代表的数学式发生了变化,直到得到满足要求的结果。GEP 算法的应用范围很广,曾被用于预测空气的 PM2.5 浓度②。在这个课题里,研究者选择了 7 个相关参考变量,如温度、相对湿度、风速、风向等,选用的运算符包括+、−、×、÷、sin、cos 等 12 种,初始的染色体种群数量为 100,每个基因的头部长度为 10,基因数目为 6。经过 120 代的进化后,得到的染色体如下:

$$*/C_3 \; +\sin X_2 \; -\log{*}\cos X_6X_4C_3X_5X_3X_1X_2X_4C_8X_2X_5 \; +$$
$$\mathrm{sqrt}{*}/X_5 \; +{*}\cos\sin X_2X_1X_3C_2C_5X_4X_1C_1X_2X_4X_3C_0X_6 \; +{*} \; +$$
$$X_3\sin X_2{*}X_1\cos X_2\exp C_1X_4X_2C_0X_2C_3X_2C_4X_2X_3C_9 \quad +\mathrm{sqrt}/$$
$${*}X_2+-X_1{*}/X_4C_0X_2X_4C_6X_1X_5C_7X_2X_4C_6++-\cos{*}\log/\sin{*}-$$
$$X_3X_4X_5C_3X_4C_0X_6C_4X_1X_2C_5 \quad +{*}\exp X_2 \quad +/{*}\sin \quad -$$
$$X_3X_4X_1C_3X_4X_2C_0X_5X_6X_2X_1C_5$$

① 唐常杰, et al. “基于基因表达式编程的知识发现——沿革,成果和发展方向.” 计算机应用 24.10 (2004): 7-10。

② 刘小生,李胜,赵相博. “基于基因表达式编程的 PM2.5 浓度预测模型研究.” 江西理工大学学报 34.5 (2013): 1-5。

该染色体对应的数学式为

$$y=-0.232\,055*[X_2-0.232\,055*X_4-\cos X_5]/\sin(\log X_6)+\mathrm{sqrt}[X_5*(\cos X_3+\sin(-0.576\,05))/(X_2*X_1)]+[\sin(X_1*\cos X_2)+X_2)*X_3+\mathrm{sqrt}[(X_1+0.269\,745*X_2)*(-X_4/0.650\,543-X_4)/X_2]+[-0.266\,113*X_4*\sin X_5-\log(0.336\,273-X_5)+\cos(X_3/X_4)]+X_2*\exp[(\sin X_1)/(0.824\,242-X_4)+X_3*X_4]$$

使用算法自行计算出来的结果,对空气中PM2.5浓度进行预测的结果,准确度明显高于其他预测模型,预测精度非常高。

GEP算法可不是就能预测一下PM2.5浓度这么简单,对于时间序列的数据,如气象、地震、太阳黑子等,甚至对于股票价格,都有很好的预测效果。此外,在因式分解、谓词关联规则挖掘、微分方程求解等很多方面都有丰富的应用。不知道大家有没有注意到这样一个事实,算法自行推导出的结果,对我们来说,往往难以理解,也很难给出一个直观的解释(比如上面的PM2.5浓度预测的式子)。这是不是可以理解为算法(或机器)的思维与人类不同?人们正在逐渐依靠能够自行推演出模型(结果)的算法和机器,但这些模型却超出人类的理解范畴,以一种异于人类的方式来“思考”这个世界。

人类在认识世界的时候,往往需要借助模型的帮助。模型,是对于真实世界中的事物、现象、过程或系统的简化描述,或是其部分属性的模仿。如果需要了解和认识的客体较为复杂,超过了人类处理能力的限度,那么人们常常采用的方法是,忽略或简化那些他们自己认为不重要的、影响轻微的因素,主要考虑那些具有明显影响的因素;还不行的话,则对客体进行分解或割裂,化整为零,分别进行考虑。通过这样的处理,就可以对客体构建实体模型,或是抽象的数学模型,进而对客体建立起直观、深刻的认识。通过模型来认识客体,是一种可行的办法。然而,通过简化或分解处理得到的模型,是否还能准确反映真实的客观世界?

人类历史中出现过很多有名的模型，例如在战争中经常用到的沙盘模型、金融机构中的各种借贷模型、马尔可夫模型，还有小朋友玩的乐高模型等等。BackChannel 网站在一篇著名的文章中列举了两个经典的例子来说明人类构建的模型中可能隐含的问题①。

密西西比河是北美大陆水量最大、流域面积最广的河流，滋润着美国41%的土地。1943 年，二战还未结束，出于治理需要，美国陆军工程兵部队打造了一个 200 英亩大的密西西比河及沿岸土地的模型，这个超级模型用来进行各种模拟实验，以便人们了解，如果发生水患，沿岸的城镇会面临什么状况。得益于充分的研究和准备，奥马哈市才安然度过 1952 年出现的洪灾，避免了超过 6500 万美元的经济损失。此外，有研究人员表示，当时从那些看似简陋的模拟实验中得到的结果远比现在的数字模型准确。

第二个谈到的模型来自经济学，在这个模型里，流动的水也是一个重要组成部分。

1949 年，当时还是伦敦经济学院一名学生的新西兰经济学家威廉·菲利普斯(William Phillips)，在研究英国经济运行过程时，构建了一个长、宽、高分别约为 1.2 米、1 米、2 米的模型，这个模型被命名为 MONIAC (Monetary National Income Analogue Computer，货币国民收入模拟计算机)。在 MONIAC 中，包括一些透明的管道和水箱。水箱代表英国经济系统的不同方面，比如最顶上的水箱代表英国的总财富，管道中流动着的染了色的水代表财富或金钱在系统中运行流转的过程。

MONIAC 模型可以用来展示凯恩斯经济政策带来的影响。但实际上，虽然 MONIAC 看起来很高大上，但它的准确度和可靠性还比不上密西西比

---

① Backchannel. *Our Machines Now Have Knowledge We'll Never Understand*. https://backchannel.com/our-machines-now-have-knowledge-well-never-understand-857a479dcc0e。

河模型。这里的原因可能是，MONIAC 中只包含了影响国家经济状况的重要变量，而没有包含全部变量。但在密西西比河模型中，也有很多人类无法列出的变量的影响，例如河流沿途的山脉、植被变化，或是月球引力的影响。但为什么密西西比河模型却能够得到与真实情况基本吻合的结果呢？

对比密西西比河模型和 MONIAC 模型，前者是一个物理模型，后者是一个人工模型。例如，你如果想看看密西西比河沿岸的山体塌方时，垮塌的山石滚入河道会有什么情况，你要做的只是制造一些相应比例的岩石模型，扔到密西西比河模型中，就可以直观看到发生的结果。你完全不需要花费大量的时间和精力去学习和了解流体力学的相关知识。也就是说，当模型的比例对结果没有影响时，物理模型就会给你想知道的答案。对此，高级水利专家斯坦福·吉布森（Stanford Gibson）的说法可能更通俗易懂："物理模型会自行模拟所有的过程。"

在 MONIAC 中，利用有颜色的水来模拟货币在经济系统中的流转运行过程，从而验证各种经济理论。与上面讲述的密西西比河这样的物理模型不同的是，MONIAC 的构建是基于人们事先进行的逻辑推理而形成的数学模型。这些模型中其实隐含了诸多限制，在 MONIAC 中水箱的大小、水量的多少、水管的粗细、流向控制则体现了这些限制的存在。这些限制并非自然形成，带有很深的人为印记。密西西比河模型中看似没有多少假定的限制，其实不然。密西西比河模型假定了比例无关，在真实环境中会发生的事情在 1∶10 000 的模型中也会发生。模型还假定了河流沿岸山脉的高度、沿岸田地里是否有农作物等因素对实验的结果没有影响。

两种模型中都有假定这种限制存在，那为什么密西西比河模型能取得很好的结果，而 MONIAC 得到的结论有时却离事实相差很远呢？其实，在密西西比河模型管用的情况下，人们也不明白它为什么行得通。密西西比河模型的构建没有基于任何数学模型，它的形成基于客观存在，仅仅是在比例大

小或某些不重要的因素上有所不同。或许,它能够有很好的效果就在于人们没有过多地干预,它不需要人们理解它为什么管用,让它模拟客观存在并自行演进就好,无须再施加人们的逻辑推理所带来的限制。有可能,人们根据自己的理解强加给模型的限制,正是使模型得出不准确结果的原因。

一直以来,人们都是先手工设计模型,然后再将模型通过计算机来处理和验证。这种方法的问题在于,要想得到更好的结果,人们就需要不断升级模型,以使它变得更加具体、准确。但在这个过程中,人们无法保证自己的理解的正确性,也无法保证不会引入更多干扰,这些都与人们的知识和理解紧密相关。得益于机器学习的出现,人们可以有另外的选择。人们可以尝试放手,让机器以算法思维来理解人类的世界,自行进化。

目前看来,Google 公司的 AlphaGo 项目是让机器按照自己的“意愿”进行学习和理解人类世界的一次有趣尝试。AlphaGo 在战胜了围棋大师李世石后,在非正式情况下,它的升级版本 Master 又取得连胜 60 盘的梦幻成绩。从理论上讲,围棋中可能的变化数达到了 10^{350} 之多。AlphaGo 进行决策的大脑是由多层神经网络构成,它使用了发生在人类棋手之间的 16 万盘对局约 3000 万步走法对自己进行训练,来理解怎样才能在围棋这个游戏中取得最高的胜率。支撑这一切的硬件也仅仅是 48 个 CPU(处理器),再加上额外的 8 个 GPU(图形处理器)。

尽管 AlphaGo 已经是无可非议的世界级棋手,但它没有办法讲出能够让人类棋手学习和理解的道理,因为决定其走法的核心是“概率”,这跟人类对围棋的理解完全不同。AlphaGo 的决策是依靠多层神经网络中的数十亿个连接,它自行创造了一个极其复杂的模型。这个模型中产生的巨量应变走法的目的仅仅是为了在围棋中战胜人类,并无其他。如果你非得用人类“渺小”的大脑去试图理解 AlphaGo 的每一步走法,在令人头昏脑涨的解释里一定会无数次出现“加权连接网络”,这些连接负责将结果传递到下一层神经网

络。人类的大脑根本无法记住所有的变量和权重，以至于根本无法进行基于这些变量和权重的计算。就算能够计算，人类也不知道如何根据计算结果去下围棋，这些结果与人类的认知完全不同，这是算法的认知逻辑，是内化的人类下棋时神经状态的运行原理，并不能帮助你理解它的走法，只有机器（算法）才明白这究竟是什么。

人类无法理解机器（算法）带来的知识，也有两者认知模式不同的原因。举例来说，人类识别数字“0”，是根据在学习过程中所获得的知识或规则来判断，例如，数字 0 特别类似圆圈，也特别像字母 o。但机器则不然，给机器提供成千上万个书写的数字 0 的图片，它会将这些图片转换为二进制存储的矩阵并寻找这些矩阵的相似度，进而根据这些学到的知识来判断新的图片中是否有数字 0，同时还要根据上下文来排除字母 o，说实话，识别 0 和 o 对于人类来说都有相当的难度，对于这一点，在网上填过随机图片验证码的人一定深有体会。

在算法的作用下，机器的智能化程度越来越高，在发现模式、做出结论方面的能力已经超过了人类。并且，随着硬件的高速发展，计算机在建立模型时可以做得尽可能地大，只要算法和硬件能够支持。比如，人工神经网络层数越多越准确，当然复杂度也更高，在当前的条件下，深度达几百层的神经网络，已经不是问题。模型能做得足够大，其中能够容纳的变量就更多，就可以考虑更多的影响因素，而不用通过简化或分解以迎合相对简单的模型。但是，这同时也意味着，我们想要了解的客体或现象，必须依靠机器得出的结果。然而，对于机器怎么建模、如何求解的过程，则是人们无从得知、也无法理解和解释的。对于此，机器学习专家亚当·盖特吉（Adam Geitgey）有一段形象的描述：“通过泛型算法能够告诉你一组数据的有趣之处，并且你无须针对问题编写任何代码。你只需要给泛型算法提供输入数据，然后它就会根据那些数据建立自己的逻辑。”

在不少人看来，这样的模式，有可能颠覆人们传统的认知观念。这种使用异类智能的方式，引发了人们对一直以来的西方传统中植入的假设思维产生了疑问。正如美国著名评论家大卫·温伯格（David Weinberger）在媒体网站 Backchannel 的文章里所说："以前，人们认为，知识的存在是为了简化这个复杂的世界。但当前的情况，似乎在说人们搞错了，认识我们身处的世界，可能反而需要放弃去理解它。"

人类还能做点什么

知识是确证的真信念。

——柏拉图，思想家

自从人们开始在绳子上打结，或者在木棍上刻凹痕来帮助计数以来，人们一直都是通过这个世界中客观存在的事物来辅助自己了解和认识世界。人们很清楚每个绳结或凹痕所代表的含义，或者再复杂一点，人们知道修建密西西比河模型的目的是为了应对洪水的发生。人类一直以来都有一种信念，认为这些可以带来知识的模型，本身能够准确地反映世界的运转方式。但现在正在发生的事情，却正在使人们的这个信念变得摇摇欲坠。人们越来越多地依靠那些并不符合人类推理逻辑的东西，虽然人们还没有办法去理解那些冷冰冰的没有感知的小伙伴们推导出答案的方式和过程。如果说知识的概念确实是古希腊哲学家柏拉图所定义的那样，那么人们应该如何去理解这种新得到的知识？要知道，现在这种新型知识不仅难以解释确证，更是无从解释确证。这开始令人类感到不安，人类总是对未知的事物心怀恐惧，就像在古时，人们不了解月食现象，恐惧之下，只好用"天狗吃月"这类带有神学

色彩的说法来解释一样。

计算机从一出现，就在随着硬件的发展，不断扩展其能够处理的信息量。我们可以类比前面定义的信息“人比”，给出“机器比”的概念，即计算机能接收和处理的最大信息量。通常来说，机器比大于人比。当人类世界进入赛博时代后，信息开始爆炸性增长。对人类来说，获取大量知识是困难的，因为这需要汇集一批人比的信息，而人与人之间建立联系的难度限制了这一汇集过程。与人类不同，赛博的出现使机器之间建立连接变得毫无问题，同时，机器之间的沟通远比人类更加高效和简单，当机器不断变革，机器比也被无限放大。当下，计算机能够容纳网络中的全部信息。这些信息不仅包括存储在数据库中的历史内容，还包括来自人类行为和遍布全球的各类传感器产生的实时信息。现在的信息更像是不断汇聚融合的溪流，而不是储藏在数据仓库里的资源。

从信息的产生量和存储量，以及赛博世界的连接关系拓扑，可以让我们认识到身处的世界是多么庞杂，多么不确定。几千年来，大约从人们开始理性思考开始，人们就毫不动摇地认为简单模型会反映宇宙的简单性。现在，正是人类自己创造的机器给了我们当头一棒，它告诉人们，世界在它的眼里是何等的错综复杂，以人类大脑的认知无法理解，需要借助由人类和计算机组成的网络才能认识到：这个世界完完全全是混沌的和不确定的。

面对这个混沌的世界，当人们开始依赖于不可理解的模型给出的知识，这里会出现一个问题，谁来对这些新型知识确证？几千年前，柏拉图就告诉我们，没有确证的信念不成为知识。然而，机器可以确证的，往往是人类无法理解的。一个可能的应对是人们需要放弃某些知识。事实上，在现实中，我们已经开始这么做了。例如，法院已经规定，未经合法方式（授权）所取得的信息不能成为证据，因为如果允许这类证据出现，显然会给公诉方带来收集这些证据的动机。另外，机器学习算法已经应用于社会信用评级领域。在我

国,人们熟悉的芝麻信用分就是一个典型的例子。同时,人们各自的信用分又各不相同,这是因为算法采用的评判标准包括身份特质、居住情况、履约能力、消费习惯、人脉关系等诸多因素。不过,在不同的国家或地区,一些特殊的评价因素可能不适用。例如,在美国,评级算法可能会发现,不同宗教信仰的人,信用风险会不一样。尽管这可能是事实,但美国的法律规定了这些机器得出的知识也不能用于对用户进行信用评价,信用评分公司也会被禁止使用与信仰属性有关的数据。

另外一方面,机器展现出来的能力的确非常强大,但它给出的结果一定是正确的吗?机器是从人类历史数据中进行学习,由于人性中确实多少带有一些灰暗属性,这可能会给数据中带入负面的干扰,从而给机器形成的结论带来影响。比如,在美国用来评估申请保释的罪犯带有潜在风险的系统,在训练过程完成后,开始进行实际评估操作时,人们发现它往往会为白人罪犯网开一面,但对那些犯罪记录较少非裔或亚裔美国人则异常严厉。不难解释,这是机器学习有了人类历史数据中的种族偏见,虽然美国一直自诩人权至上。

2013 年美国波士顿马拉松爆炸案后的 7 月 31 号,家住纽约纳苏县的女记者卡塔拉诺在家突然遭到 FBI“反恐联合工作组”6 名特工的搜查。据卡塔拉诺称,FBI 的特工大约花费 45 分钟时间搜查她家,盘问她丈夫的网上搜索和他到海外的公务旅行。造成这次搜查的原因居然是因为卡塔拉诺在亚马逊网站上搜索过高压锅,高压锅正好是波士顿马拉松爆炸案中嫌犯使用的工具,不知道什么原因触发了监控软件发出了疑似恐怖活动的警告。

早年 Google 的算法也出现过乌龙事件。2009 年,Google 的工程师在世界顶级学术期刊 *Nature* 上发表了一篇关于流感疫情预测的论文,同时聪明的工程师们还上线了流感预测系统 GFT(Google Flu Trends)。当时,美国国家疾控中心的预警通告要比实际情况滞后两周左右,而 GFT 得出的预警

滞后时间仅为 1 天，有时只是几个小时，亮出的成绩单非常惊艳，在当时引起巨大反响。但好景不长，到了 2014 年，有学者在另一本世界顶级学术期刊 *Science* 上发文公布了 GFT 的预测情况。从 2011 年 8 月起的 108 周时间里，有 100 周 GFT 的预测均大幅偏高，其中 2012 年 12 月的一次预警比实际情况高出近乎一倍。为什么会出现这种情况？原来 GFT 的预测原理是基于一个简单的情形：如果在某一区域的某一时段，出现大量关于流感的搜索，那么该区域可能出现了流感疫情。其实，Google 的工程师们并不知道关键词搜索与流感爆发之间是否存在关联。让工程师们没想到的是，GFT 的成功引发了人们的好奇心，人们都想来看看这个神奇的系统，结果造成了预警信息失效。

从这些例子中可以看到，人类或许能够在事后及时修正偏见，但机器在学习过程中可能会重新建立起我们当前已经修正的隐含在数据中的偏见。因此，对人类来说，也许我们需要做两件事：一方面，人们应当禁止一些确证类型，以避免造成不良的社会影响。另一方面，人们应该着手制定一些规范和监督手段，类似于阿西莫夫的机器人三定律那样，确保有效减小和避免错误的发生。

拥抱“异类”智能

我看到一个以前没有见过的轮子不停地旋转，时而上升，时而下降……命运之轮将我们多次推入深渊，但是它又多次将我们送上巅峰，如此循环往复。我们应该了解命运的转轮。

——万那·怀特（Vanna White），演员

人类创造的机器在算法和赛博的作用下,其能力正日益变得强大。它们完全不需要人们事先将需要的信息进行缩减,机器比已经无限扩大。正是由于它们拥有了这种新能力,人们正逐渐习惯于交给它们所有可能需要的信息,然后再问自己想知道的问题。但在背后,人们无法理解它们的思维方式,就像外行可能无法完全理解 TCP/IP 协议的细节。而且,基于计算机的确证,在本质上完全不同于人类的确证方式。机器的智能是一种“异类”智能,不过,异类并不意味机器的认知是错误,而仅仅是指确证方式与人类不同。要说到理解认知客观世界,真相可能会令人类感到有点沮丧,那就是机器比我们人类在任何时候都要更加接近世界的本质。

在获取知识这件事情上,一直以来人类都是利用工具来完成的。几千年前的放牧人,由于不懂数学,他需要准备一些小圆石子,以确保放牧归来时牲畜的数量与出发时一致,至少不会变少;现在的学者在思考问题时,很可能还需要稿纸或白板来做推导的工作;建筑师需要够大的纸张、直尺、铅笔甚至三维模型来思考建筑的构造。现在,这些领域的从业人员都已经转为使用电脑了,前面谈到的放牧人甚至可以通过给牛羊身上加装微型传感器,然后通过最新的手机 App 来管理自己的牧群。然而,情况还是没变:我们仍然在利用工具来理解这个世界。只不过,上述这些利用工具的方式,是人们能够理解的方式。

而人工智能和机器学习的发展,则进一步凸显了人类的理解力相对于其给自己设定的任务的不足。人类仍然在利用工具理解世界,只是现在的工具换成了计算机。当机器使用神经网络等人类所无法理解的确证方式给出得到的知识时,人类往往手足无措。人们可以通过强调 AlphaGo 不断战胜人类棋手,以及算法驾驶的汽车确实更少发生交通事故等事例,来证明机器通过学习得出的结果很有可能是知识,但人们还是不能理解 AlphaGo 为什么会下这一步,而不是那一步,也不能够理解明明应该往左转,但机器却控制汽

车往右转。这里面涉及各类信息输入、机器的决策过程,正如在前面反复讲到的,这些东西即便是最聪明的人脑也无法理解。

在 *Our Machines Now Have Knowledge We'll Never Understand* 中提到,人类其实一直存在一个幻觉:只要计算机还能够通过模型把人们的想法具体展现出来,人们就还是认为世界在按照其拥有的知识(模型)所理解的方式运转。然而,一旦计算机开始按自己的方式创建模型,并且这些模型超出人们的理解范畴,人们就失去了那种令人心安的假设。我们自己创造的机器让人类知识论的局限性变得显而易见。当我们不能理解计算机创造的模型时,我们还能心安理得的认为计算机一定是人类的奴仆吗?

科技在人类发展历史的长河中总是扮演着最活跃最革命的角色,它的发展是不可阻挡的。而面对发展的未知,我们难免恐惧。机器能做到什么程度?智能的爆炸是什么样的?机器学习革命会以什么方式呈现?算法如何运行?我们需要进行怎样的认知?也许只有当我们能够真实地去面对这些问题,划定它的边界和围栏,而不是回避它、否定它和阻止它的时候,我们才能真正地迎来这个可以被计算的世界,一个由算法定义的世界。

《算法统治世界——智能经济的隐形秩序》思维导图



算法统治世界——智能经济的隐形秩序

- 绪论
- 第1章 算法到底是什么
  - 如何定义算法
  - 算法的载体——图灵和图灵机
  - 从图灵测试到人工智能
  - 算法的复杂性
- 第2章 共享经济该如何共分利益
  - 公平分配从来就是一道难题
  - 功利主义的分配方案——Shapley值
  - 平均主义的分配方案——核
  - 没有绝对的公平
  - 共享经济平台如何设计最优分配方案
- 第3章 匹配算法——双向选择市场里的丘比特
  - 如何在网上找到一个靠谱的女朋友
  - 价格在赛博新经济市场的失灵
  - Gale-Shapley匹配算法：一种更为稳定的匹配设计
  - 匈牙利算法：一只最大匹配的丘比特之箭
  - 如何从多对一的匹配中获得最佳选择
- 第4章 动态定价，应对供需剧烈变化的赛博新市场
  - 动态定价不仅仅是一系列传统定价方法的简单组合
  - 动态定价在新经济领域的应用
- 第5章 算法，让数据有了价值
  - 算法，点石成金的力量
  - 大数据算法的新征程
- 第6章 你的数据究竟该卖多少钱
  - 数据该如何买卖
  - 给数据定价是个技术活
  - 破解数据交易的不可能三角
- 第7章 谁来保护我们的隐私
  - 我们还有隐私吗
  - 在混战中寻求算法保护
  - 被算法设计的隐私保护
  - 共建数据隐私新秩序
- 第8章 信任的基础——区块链
  - 拜占庭将军问题
  - 拜占庭将军协议及容错系统
  - 共识机制先驱——P2P
  - 区块链的原型：比特币
  - 区块链：让信任成为一种社会共识
  - 区块链技术：驱动赛博经济进入智能经济时代
- 第9章 未来：赛博智能经济